SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)
Vol. 53 No. 04 (Dec-2021)

Analysis and Comparative Study of POS Tagging Techniques for National (Urdu) Language and other Regional Languages of Pakistan

Rahmat Ali Rajper, Samina Rajper, Abdullah Maitlo, Ghulam Nabi

Department of Computer Science, Shah Abdul Latif University Khairpur Mir's, Sindh, Pakistan

Article history
Submitted
March 2021
Reviewed
Sep. 2021
Accepted
Nov. 2021
Published
online
Dec. 2021

Abstract

Defining algorithms and techniques to enable computers to understand human language is the Natural Language Processing (NLP), which is an integral part of speech recognition. Parts of Speech (POS) is considered as one of the well understood problems of Natural Language Processing, in which natural language words and sentence are tagged or assigned grammatical classes, because tagging a single word by human hand is a time consuming and tedious job. To automate the tagging job is the way to automate the lexicons of the text of a language. Many of the languages are enriched with their POS tagging systems. Pakistani regional languages are less developed due to the many reasons and much of the work is needed in POS tagging system. Some of the regional languages have their POS tagging systems but still they need some more attention to refine their system. Some of the languages need to develop from the scratch. Balochi language has no any POS tagging system. This study presents the comparative analysis of POS tagging approaches for the national language (Urdu) and other regional languages of Pakistan. The approaches and their data sets used and their reported results are presented here.

Keywords: NLP, POS, Speech

Introduction

Enabling computers to behave like human being is a branch of computer sciences as Artificial Intelligence. Defining algorithms and techniques to enable computers to understand human language is the Natural Language Processing (NLP), which is an integral part of speech recognition [1]. Parts of Speech (POS) is considered as one of the well understood problems of Natural Language Processing in which natural language words and sentence are tagged or assigned grammatical classes because tagging single word by human hand is a time consuming and tedious job. To automate the tagging job is the way to automate the lexicons of the text of a language. The process of tagging starts with getting a selected word or sentence of a language and assigns a POS tag to every single candidate word and then the new output text is generated along with the tagged data [2]. Significant efforts have been made for the POS tagging of western languages such as German, English, Indian languages Tamil, Telugu, Bangla and others but very small work has been done for regional languages of Pakistan. Some studies have been found on Sindhi [3] [4], Pashto [13], Punjabi [16] [18], and Urdu [12] [14] [15] [23]. This study focuses on the comparison analysis of various approaches for the tagging of Pakistani regional languages. The analysis will pave the way for the selection of suitable technique to be used for these language tagging. This study has compared various approaches used by western and south Asian languages and suggest the suitable approach to be applied for Pakistani national and regional languages.

Cite this:
Rajper RA, S. Rajper, A. Mailto, and G. Nabi (2021). Analysis and comparative study of pos tagging techniques for national (Urdu) language and other regional languages of Pakistan. Sindh Uni. Res.J.-SS 53:4 44-53

Corresponding author
rajperghulamnabi@gmail.com

The discussion is presented with the conclusion of the techniques to be improved and enhanced and applied to regional languages of Pakistan.

Various tagging approaches has been used for POS tagging systems of Pakistani languages, the identification of more suitable technique for POS system of Arabic script based languages is different. Hence in this study, POS tagging techniques will be carefully reviewed and analyzed and appropriate technique will be suggested on the basis of reported results.

This study is limited to regional languages of Pakistan including the National language of Pakistan including Sindhi, Pashto and Punjabi. Pakistan is one of the populated country with large number of speakers containing more number of languages.

The outcome of current study is a comprehensive comparative analysis and some of the suitable recommendations to utilize an existing approach intact or there is a need of improvement or fine tuning of the approaches for the use of regional languages.

The various languages spoken in Pakistan will be presented followed by the some of the other languages and their Parts of Speech tagging systems. Techniques covers some of the POS studies based on various approaches incorporated for the development of various POS tagging systems of languages.

POS tagging for Various Regional Languages of Pakistan

Sindhi POS Tagging

Word net has been applied for the tagging system of Sindhi Parts of Speech (POS). The study highlighted the characteristics of Sindhi languages pertaining to POS tagging system such as the lexical and morphological ambiguity. Various algorithms have been proposed for Sindhi tagging system such as disambiguation rules for Sindhi words, tagging of Sindhi words and tokenization of Sindhi words.

The results have been presented by applying WordNet and without WordNet and an overall accuracy has been reported as 96.28% without net and 97.14 with word net. The results have been presented with training, testing corpus and unknown words [3]. A morphological analyzer is proposed for Sindhi language by [4].

Urdu POS Tagging

A POS tagging system has been created for the limited resource scenario contacting various approaches along with morphological features to develop a POS tagger for Urdu.

In [15], a transformation based POS tagging system for Urdu language is presented using error driven learning. For the solution of disambiguation problem in Urdu, a data driven technique Brill's transformation-based learning has been (TBL) used. The TBL approaches derives rules automatically from the corpus and produces more accuracy as compared to other approaches and provides more advantages than any other approach used for tagging systems.

In [23], the new Urdu POS tag set design schema is presented. The overall system accuracy is reported as 96.8%. The corpus has been divided into 20% testing and 80% of training selecting file at random.

Pashto POS Tagging

In [13], a Pashto parts of speech tagger based on rule base approach. The study claims as a first ever attempt of such kind of research on rule based approach for Pashto tagging system and the parser which have been used for Pashto language. The POS tagging algorithm takes input of text which tokenizes the given text. The tokens are searched in lexicons then tokens are marked and all the tokens are tagged and rules are described so that multiple tags cannot exist resultantly the tagged output text is produced.

The first experiment contains 100 number of words with applying 10 rules which produce the 40% of accuracy. Various experiments have been proposed with varying number of rules such as 10, 40, 70 and 120. The maximum number of rules were applied on 100,000 words in lexicon producing 88% tagger accuracy.

Punjabi POS Tagging

A work on Punjabi language is presented by Kaur et al., (2015) for the tagger of Punjabi language applying reduced tag set. The study identifies the problem of sparseness in a previous study of [18] due to the large data set of 630 tags available in their study. The identified problem of sparseness has been dealt in their study by applying 36 tags. The technique has been used to improve the tagging mechanism is suggested by the Technical Development of Indian Language (TDIL). Their system starts with the raw corpus collection of Punjabi language followed by the tagging of the collected corpus either manually or available existing tagger. The experimental results show that the precision of every corpus of five types is 100% whereas recall rate is from 85.2% to 99.6% [18].

Indian Languages

A comprehensive survey and comparison of tagging systems have been presented in [19]. The study presents the basic building blocks and architecture of POS tagger and other terms such as tokenization,

ambiguity lookup, corpus, tag set, WordNet and others. The various approaches have been surveyed along with their results such as rule based, stochastic and hybrid.

Malayalam

A post tagging system for Malayalam language is presented for the question and answer system. The approaches applied are POS tagging analysis and Vibhakthi. For the creation of relation or donation with a verb, morphophonemic notations are used for noun inside the sentences of Malayalam language.

South Asian Languages

A remarkable work has been done on various languages but a tagger for south Asian languages has been proposed in [25]. A 26 tag set has been used for all three selected languages. The approaches selected were n-grams, HMM, Bigram and Unigram. The results for Bangla corpus while experimenting on HMM was 63.6, Unigram 56.9, Bigram 55.5 and Brill 69.6 where clearly Brill outperformed all other taggers. For Hindi Brill also outperformed other approaches with 71.5%.

POS Tagging Techniques

Various techniques of tagging system for Bangla system has been presented in [20]. Various approaches have been tested for performance and the tagging language set is Bangla. The approaches were selected from statistical and transformation based approaches. The statistical approaches were selected were HMM and n-gram. From the transformation based approach the Brills tagger was selected. A very limited set of annotated corpus was selected and the reason has been claimed as the system is at initial level.

The main theme of the study is to identify the best performer from these selected techniques while having a limited resource. The English language has been used to verify the approaches and conclude and prove the behavior and performance of the selected approaches in case when there is a substantial amount of annotated corpus. The NLTK has been used for the experiments of unigram and bigram for the testing the approaches. The two tag sets were tested and used for the experimentation. For Bangla a 41 tag set has been used [21] and for English Brown tag set [22] has been used.

The proposed tag set has two levels including a simple or basic level containing 12 tags and the other level fine grained which contains 41 tags. Most of the experiments were performed on 41 tag set whereas some of the experiments were also carried out with level 1, a 12 tag set. A table containing number of tokens along with the accuracy achieved by various approaches have been reported. For the 12 tag set the highest accuracy shown by HMM was the 49.4% when 3016 tokens were used for experimentation. For the same 12 tag set the highest accuracy of Unigram is reported as 71.2% with 4484 tokens.

The Brill accuracy is 71.3% reported with 4484 tokens on the same 12 tag set data. For the set of 85 sentences and 1000 tokens the tag set of level 2 with 41 tags the accuracy of HMM is reported 46.9%. The Unigram accuracy is reported 42.2% as maximum with 4484 tokens. The maximum accuracy for Brill approach is reported as 54.9%. The performance of selected approaches has also been tested on 22 sentences and 1008 tokens from Brown corpus and the Maximum accuracy achieved by HMM is 87.8% with 90000 tokens, Unigram 78.9% and Brill accuracy 83.4% with 100057 token [20].

Methodology

In this section we will discuss the approaches and theories behind working of many HMM based and N-gram approaches. Some of the comparison analysis will be presented here. Many of the research has been done on languages like English and some of the other languages such as German, Latin and Indian languages. Many of the other languages have POS taggers and corpora of these languages have also been created for the testing of these POS tagging system. Pakistani languages are rich in literature but there is a very small amount of study has been carried out for these regional languages in context of POS tagging system. The focused languages are these Pakistani languages.

Corpora

Training is the important aspect of POS tagging system; a good training definitely produce more accurate results. For the purpose of training a well annotated corpus. The job of annotation can be performed on any one stage of the three namely phrase or clause level, POS level and dependency level. A common corpus for English is the NLTK corpus [27]. Various languages use their prebuilt corpora or the new corpora is being built for their testing and training purpose. In this study we present the various aspects of database or corpora used by regional languages of Pakistan.

As discussed earlier that there is a very small work done on regional languages of Pakistan. One of the regional languages is Balochi and to the best of our knowledge there is no any POS tagging system available for this language. The available and existing POS tagging systems are analyzed here and the recommended approaches are presented here. An in-depth analysis of the available systems of the POS tagging systems of regional languages of Pakistan

were carried out so that the problems and issues can be understood. Then the reported results are compared so that the results can be compared with some sort of standardized way.

Taggers: Various types of taggers are available produced by researchers around the world. Some of the taggers are also available commercially. Unigram and bigram are the most popular taggers available. HMM and brill tagger are the other names of taggers. In case of small data all taggers may produce the same results.

Unigram Taggers

A statistical algorithm in which a single tag is assigned to each token. As the name implies the unigram means one for each where n=1 for tokens. The word frequent has been assigned 'adj' more than it is used as a verb. A unigram must be trained on a training corpus as it must be trained before its use as tagging. The approaches used for regional languages of Pakistan have been analyzed to check that what n gram technique they are using. The unigram tagger passes every word which is not available to training data.

Bigram Taggers

The bigram is performing as the same way as the unigram tagger with a one difference that the bigram takes consideration of context while tagging to a current word. The different context is set for each word while illustrating the frequencies. This type of frequency is distributed because of the time of the training. The context can be understood as a word to be tagged and the previous word tag. The frequency distribution plays a vital role in tagging of the context as the word with maximum frequency is given the context. In case when tagger does not find a learnt data or it encounters a context without learnt data then it turns back automatically to unigram tagger.

HMM

HMM taggers are considered as simple but HMM taggers are a little bit different where they tag a sentence at a time. It finds a most likely sequence of words or a single word. In any sentence the sequence for the tagging is selected which can increase the chance or maximize the condition defined in following formula.

P (word | tag) * P (tag | previous n tags)

The HMM technique is considered dissimilar to other tagging approaches because it selects the best combinations of tags so that the sequence of words can be tagged. Many of the sequences are selected for tagging and the best combination is selected for tagging while other approaches tag one word at a time or word-by-word gradually without thinking of the optimal combination [28].

Brill's Tagger

Brill tagger is based on transformation which is used in case when stochastic taggers fail. Brill tagger is considered efficient because it uses very small fractional space for the nth-order stochastic order. The stochastic taggers are considered as faster taggers along with high accurate once they are trained on the corpus [29] [30]. The drawback of these stochastic taggers is the size, where these approaches create so many as well large tables when back off the nth order tagging. These tables contain entries of million and very large sparse arrays which make stochastic taggers a bad choice for the use in mobile computing devices as the mobile devices are relatively lack computing power and sufficient storage. The shortage of storage and computing power are the reasons where transformation-based taggers are choice to use.

Results and Discussions

The comparison of various studies especially National language and other languages of Pakistan. Various POS tagging models have been proposed for English and other western languages. Many of the languages have their own POS tagging systems whereas very small research has been done on regional languages. More than 90% accuracy has been achieved by English and other languages.

Comparison of regional Languages

To the best of our knowledge, very small work is available for some of the regional languages whereas a considerable work is available for the Urdu language as it is a national language. The language like Sindhi, Punjabi and Pashto do possess their POS tagging systems. To the best of our knowledge there is no work found on Balochi language as it lacks a POS system as well as other computing resources. The comparison of available POS tagging systems are presented in following sections.

Sindhi POS Tagging System

A study presented in [3] applied POS tagging for Sindhi language. The study highlighted the characteristics of Sindhi languages pertaining to POS tagging system such as the lexical and morphological ambiguity. The study presents semantic POS system based on rule-based system using WordNet so that the analogical relation of text and words can be identified. The structures of WordNet have been used for the tagging tasks a popular Sindhi dictionary titled

comprehensive Sindhi Dictionary has been used for corpus collection. The selected corpus was selected on the basis of local people or speakers using vocabulary in their daily life in recent times. The results have been presented by applying WordNet and without WordNet and an overall accuracy has been reported as 96.28% without WordNet and 97.14% with WordNet [3].

Figure 1 illustrates the calculated accuracy of Sindhi parts of speech tagging accuracy without WordNet where conjunction and interjections are producing 100% results. Prepositions have produced the lowest accuracy with a score of 87.5%. The values are without application of the WordNet.

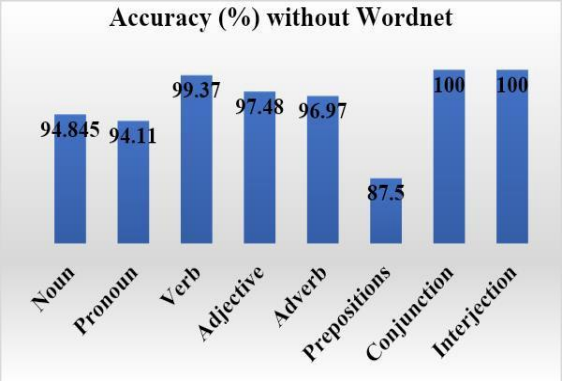

Fig 1. Calculated accuracy of word frequency without word net for Sindhi

Following **Figure 2** shows the accuracy of Urdu parts of speech by using WordNet. **Figure 2** illustrates the accuracy of word frequency by using WordNet which definitely increases the overall accuracy hence it can produce improved results. The difference can easily see with an increase in verb from 99.37% to 99.77% and adverb from 96.97% to 98.6%.

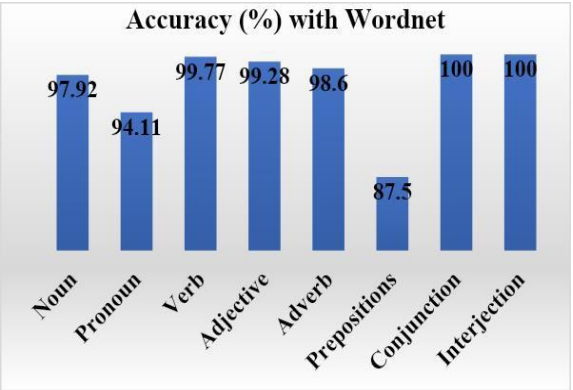

Fig 2. Calculated accuracy of word frequency with WordNet for POS

There is no any difference on preposition as it produced same results when using WordNet or not using WordNet. The combined word accuracy is illustrated in **Figure 3**. The figure shows the difference between using WordNet and without using WordNet. Both reported values are plotted together to analyze the difference clearly. The study [3] also presents the precision and recall rate for Sindhi where precision rates for known tokens and overall tokens.

Another study [4] a morphological analyzer is proposed for Sindhi language in [4]. The finite state model is a free from, open source and works on Sindhi language. For the development of this finite state transducer, an Apetium's toolbox has been used and the paradigm approach has been used for the development. The experiments were performed on freely available source of Sindhi corpus, Sindhi Wikipedia.

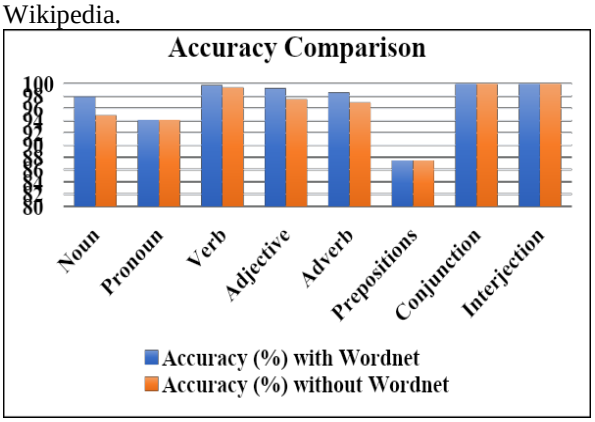

Fig 3. Accuracy Comparison (With and Without WordNet) for Sindhi.

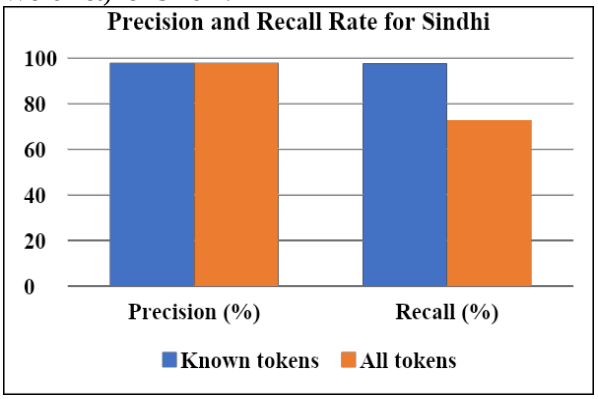

Fig 4. Calculated accuracy of precision and Recall rate for Sindhi

Another source for the experiments was the Sindhi Grammatical Framework (GF) library to define and verify the format of paradigms. The words were added manually and the corpus has been parsed so that the word lists can be created which were sorted in descending order along with frequency. The

evaluation was a twofold step. The first is the calculation of mean ambiguity and naive coverage of freely available corpus. The other step was the calculation of recall and precision. The Precision for known tokens and all tokens were reported as precision same as 97.68%. The recall rate for known tokens were 97.52% and all tokens were 72.61%. The coverage rate for wiki was 81.12%, for blogs 76.68% and as average 78.90% [4].

Urdu POS Tagging System

In [5], Anwar presented a study based on solution provided by Hidden Markov Model (HMM) for the problem of Urdu tagging system. The selected HMM model is a combined result of transitional and lexical probabilities. Various smoothing approaches have been combined to form an HMM model-based tagger so that the sparseness issue can be resolved. Analysis of Variance has been used for the evaluation of HMM based model and presented as various smoothing approaches along with the achieved word level accuracy to present the significance of results.

The most tagging error occurrences have been shown in confusion matrix. The results present the overall accuracy of implemented approaches, known word accuracy, Recall and F-Measures. The study has been claimed as a directional path of Urdu processing as such tagger system should be considered as the milestones for Urdu processing [5].

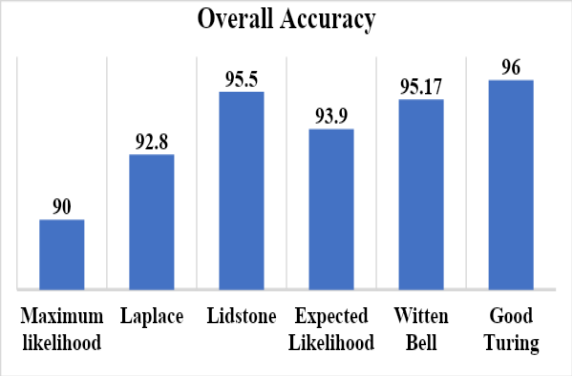

Fig 5. Calculated accuracy for Urdu

In **figure 5**, the calculated accuracy for Urdu approaches where the main approaches for the POS tagging has been shown. Various approaches have been selected for analysis and the optimal result is produced by Good Turing. The lowest accuracy is produced by maximum likelihood with the accuracy parentage of 90. A sum total of six approaches have been selected for analysis.

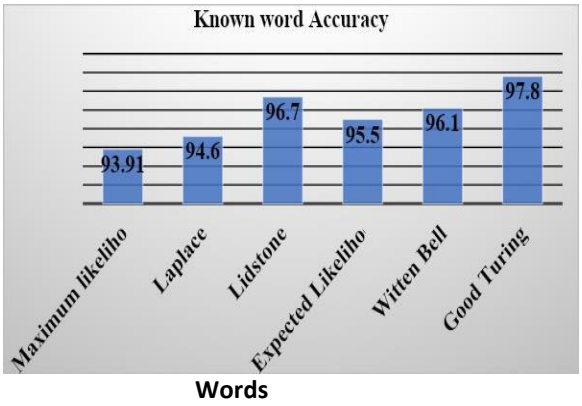

Fig 6. Calculated known word accuracy rate for Urdu

Figure 6 presents the known word accuracy of Urdu language reported by [5]. The highest known word accuracy is reported for Good Turing whereas the lowest known word accuracy is reported as 93.91%. The overall accuracy for known word accuracy is above 90% which is a better sign for the overall systems.

Figure 7 presents the calculated recall rate for Urdu language and the **Figure 8** presents F-Measure for Urdu Language. An overall accuracy of various approaches under observations have been shown in **Figure 9**. The technique Good Turing outperforms all other approaches.

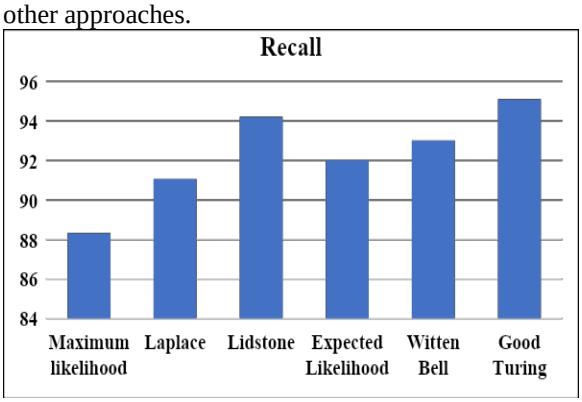

Fig 7. Calculated Recall rate for Urdu

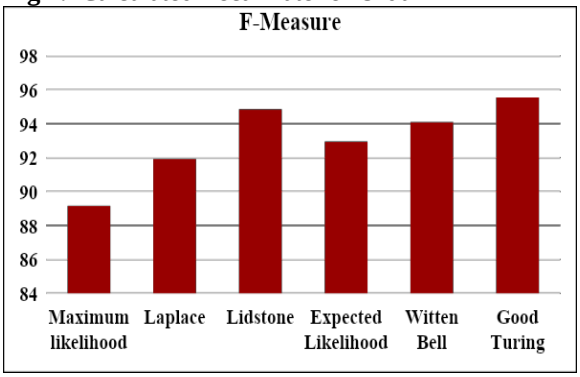

Fig 8. Calculated F-Measure for Urdu

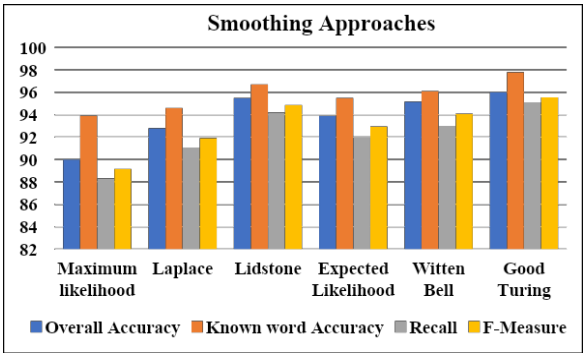

Fig 9. Calculated accuracy of smoothing approaches for Urdu.

Pashto POS Tagging System

A Pashto parts of speech tagger based on rule base approach [7]. The study claims as a first ever attempt of used for Pashto language. The POS tagging algorithm takes input of text which tokenizes the given text. The tokens are searched in lexicons then tokens are marked and all the tokens are tagged and rules are described so that multiple tags cannot exist resultantly the tagged output text is produced. The reported results for Pashto language POS tagger is shown in **Fig 10**.

Punjabi POS Tagging System

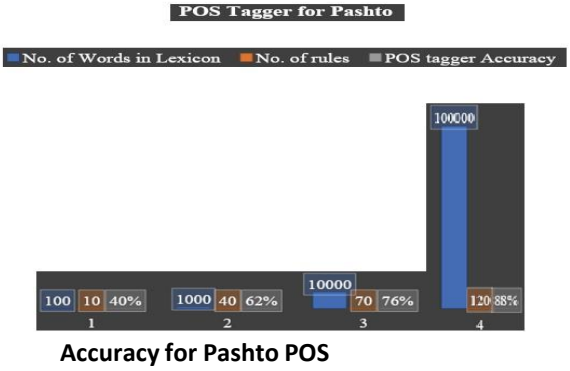

Fig 10. Reported accuracy of smoothing approaches for Pashto

[8] is a Punjabi Part of Speech tagging system based on N-gram Model. The existing system based on rule base uses only hand-written rules and fails to resolve the issues of ambiguity in number of words. A bi-gram model has been used to resolve the part of speech tagging problem. A corpus with annotation has been used to train the corpus and the bi-gram probabilities estimation.

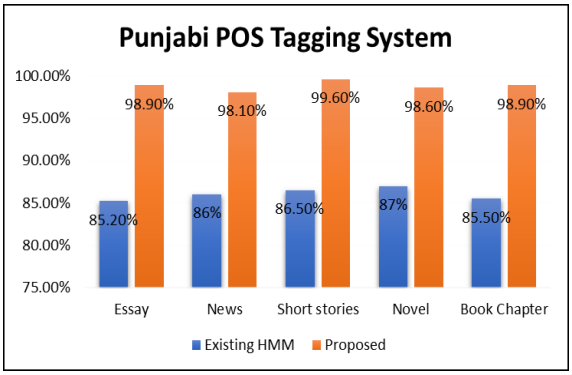

Fig. 11. Calculated accuracy of POS tagging for Punjabi

The experimental results have been reported as accuracy percentage and two sets for the testing purpose were selected. Set 1 contains 5995 number of words whereas set 2 comprise of 4007 words. The Results presented with 5233 correct tags in set 1 and 3369 in second set. The overall accuracy while ignoring the unknown words have been shown as 92.16% on a sum total of 9333 know words [8].

Balochi POS Tagging System

There has been no any work found on Balochi language (To the best of our knowledge). So, there is a gap and scope of POS tagging system for Balochi language as it is one of the Regional languages of Pakistan.

Optimal output for Sindhi and Urdu

Various approaches have been analyzed and the optimal selection can be understood a trivial job when so many options are available.

Reported Results for Various Languages

Many of the languages possess POS tagging systems and various approaches have been stated and range of results have been reported. **Fig 12** shows the reported results of various languages by researchers around the world. The maximum average reported accuracy is by Thai POS tagging systems with 99.10%.

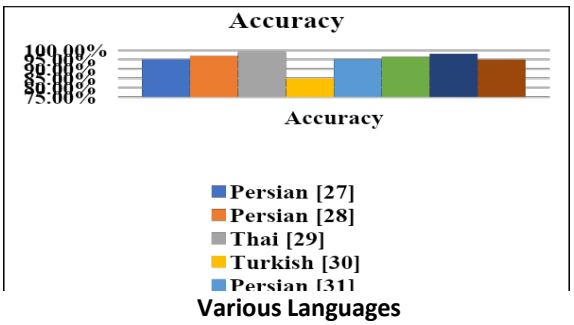

Fig 12. Reported accuracies for various languages

Discussion

Urdu (National) language and Sindhi, Punjabi, Pashto and Balochi are considered the regional languages of Pakistan. Many of the studies found on Urdu and Sindhi and a little bit work has been found on Punjabi language. Pashto tagging system found on rule based tagging with tokenizing, so that multiple tags cannot exist resultantly the tagged output text is produced. None of the work has been found on Balochi language. Research studies found on Sindhi and Urdu are very limited so it is very hard to select and apply the right technique for the development of multiscript POS tagger.

Conclusion

Many of the techniques along with their accuracies were analyzed and compared in this study and presented analysis of few approaches have been employed for the development of POS tagging system for regional languages of Pakistan. The accuracies shown in this study present over 90% accuracy of Sindhi and Urdu POS Tagging systems. Very small work has been found on Punjabi and Pashto was tagged by rule based and tagged by tokenizing so that multiple tokens were not tagged from the resultant text and nearly non-existent for Balochi.

References

[1]. Besacier, L.; Barnard, E.; Karpov, A. & Schultz, T. (2014), 'Automatic speech recognition for under resourced languages: A survey ', Speech Communication 56, 85 - 100.

[2]. Manjit, K., Mehak, A., Sanjev, K., S., (January 2015),"Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set", International Journal of Computer Applications & Information Technology, Vol. 7, Issue II.

[3]. Mahar, J. A., & Memon, G. Q. (2010). Sindhi part of speech tagging system using WordNet. International Journal of Computer Theory and Engineering, 2(4), 538

[4]. Motlani, R., Tyers, F. M., & Sharma, D. M. (2016). A Finite-State Morphological Analyser for Sindhi. In LREC.

[5]. Anwar, W., Wang, X., Li, L., & Wand, X. (2007). Hidden Markova model based part of speech tagger for Urdu. Information Technology Journal, 6(8), 1190-1198.

[6]. Khanam, M. H., & Murthy, K. M. (2014). Part-of-Speech Tagging of Urdu in Limited Resources Scenario. International Journal on Recent and Innovation Trends in Computing and Communication, 2(10), 3280-3285.

[7]. Rabbi, I., Khan, A. M., & Ali, R. (2009). Rule-based part of speech tagging for Pashto language. In Conference on Language and Technology, Lahore, Pakistan.

[8]. Mittal, S., Sethi, N. S., & Sharma, S. K. (2014). Part of Speech Tagging of Punjabi Language using N Gram Model. International Journal of Computer Applications, 100(19).

[9]. Kaur, M., Aggerwal, M., & Sharma, S. K. (2014). Improving Punjabi Part of Speech Tagger by Using Reduced Tag Set. International Journal of Computer Applications & Information Technology, 7(2), 142.

[10]. Ahmed, Raju S.B, Chandrasekhar Pammi V. S., Prasad M.K (2002), "Application of multilayer perceptron network for tagging parts-of-speech", Proceedings of the Language Engineering Conference, IEEE..

[11]. Hasan, M. F., Uz Zaman, N., & Khan, M. (2007). Comparison of Unigram, Bigram, HMM and Brill's POS tagging approaches for some South Asian languages.

[12]. Anwar, W., Wang, X., Li, L., & Wand, X. (2007). Hidden markov model based part of speech tagger for Urdu. Information Technology Journal, 6(8), 1190-1198.

[13]. Parikh, A. (2009). Part-of-speech tagging using neural network. Proceedings of ICON.

[14]. Plank, B., Søgaard, A., & Goldberg, Y. (2016). Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss. arXiv preprint arXiv:1604.05529.

[15]. Qin, L. POS tagging of Chinese Buddhist texts using Recurrent Neural Networks.

[16]. Zheng, X., Chen, H., & Xu, T. (2013, October). Deep Learning for Chinese Word Segmentation and POS Tagging. In EMNLP (pp. 647-657).

[17]. Mahar, J. A., & Memon, G. Q. (2010, February). Rule based part of speech tagging of sindhi language. In Signal Acquisition and Processing, 2010. ICSAP'10. International Conference on (pp. 101-106). IEEE.

[18]. Motlani, R., Lalwani, H., Shrivastava, M., & Sharma, D. M. (2015). Developing part-of-speech tagger for a resource poor language: Sindhi. In Proceedings of the 7th Language and Technology Conference (LTC 2015), Poznan, Poland.

[19]. Motlani, R., Tyers, F. M., & Sharma, D. M. (2016). A Finite-State Morphological Analyser for Sindhi. In LREC.

[20]. Viswes wariah, K., Chenthamarakshan, V., & Kambhatla, N. (2010, August). Urdu and Hindi:

Translation and sharing of linguistic resources. In Proceedings of the 23rd International Conference on Computational Linguistics: Posters (pp. 1283-1291). Association for Computational Linguistics.

[21]. Mahar, J. A., & Memon, G. Q. (2011). Probabilistic Analysis of Sindhi Word Prediction using N-Grams. Australian Journal of Basic and Applied Sciences, 5(5), 1137-1143.

[22]. Mahar, J. A., Shaikh, H., & Sangi, A. R. (2011). Comparative analysis of rule based syntactic and semantic sindhi parts of speech tagging systems. International Journal of Academic Research, 3(5).

[23]. Jawaid, B., Kamran, A., & Bojar, O. (2014, May). A Tagged Corpus and a Tagger for Urdu. In LREC (pp. 2938-2943).

[24]. Rakholia, R. M., & Saini, J. R. (2015, March). The design and implementation of diacritic extraction technique for Gujarati written script using Unicode Transformation Format. In Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on (pp. 1-6). IEEE.

[25]. Bhatti, Z., Ismaili, I. A., Hakro, D. N., & Waqas, A. (2014). Unicode Based Bilingual Sindhi-English Pictorial Dictionary for Children. American Journal of Software Engineering, 2(1), 1-7. [26]. Mahar, J. A., Memon, G. Q., & Danwar, S. H. (2011). Algorithms for sindhi word segmentation using lexicon-driven approach. International journal of academic research, 3(3).

[27]. Azimizadeh, A., Arab, M. M., Quchani, S. R., (2008), "Persian Part of Speech Tagger Based on Hidden Markov Model", 9th JADT.

[28]. Shamsfard, M., Fadaee, H., (2008), "A Hybrid Morphology-Based POS Tagger for Persian", Proceeding of the 6th International Language Sources and Evaluation, pp. 3453-3460.

[29]. Ma, Q., Murata, M., Uchimoto, K., Isahara, H., (2000), "Hybride Neuro and Rule-Based Part of Speech Taggers", International Conference on Computation Linguistics", pp. 509-515.

[30 Altunyurt, L., Orhan, Z., Gungor, T., (2007), "Towards Combining Rule-Based and Statistical Part of Speech Tagging in Agglutinative Languages", Computer Engineering Vol 1. pp. 66-69.

[21 Mohtarami, M., Amiri, H., Oroumchain, F., Rahgozar, M., (2008), "Using Heuristic Rules to Improve Persian Part of Speech Tagging Accuracy", 6th Int. Conference on Informatics and Systems (INFOS2008), pp. 34-38.

[32]. Ratnaparkhi, (1996), "A Maximum Entropy Model for Part of Speech Tagging", In Proc. of the Conference on Empirical Methods in Natural Language Processing, University of Pennsylvania.

[33]. Habash, N., Rambow. O., (2005),"Arabic tokenization, part-ofspeech tagging and morphological disambiguation in one fell swoop". Proceedings of the 43rd Annual Meeting of the ACL, pp. 573- 580.

[34]. Anwar, W., Wang, X., Luli, Wang, X., (2007), "Hidden Markov Model Based Part of Speech Tagger for Urdu", Information Technology Journal 6(8), pp. 1190-1198.

[35]. Archana, S. M., Vahab, N., Thankappan, R., & Raseek, C. (2016). A Rule Based Question Answering System in Malayalam Corpus Using Vibhakthi and POS Tag Analysis. Procedia Technology, 24, 1534-1541.

[36]. Sikdar, U. K., Ekbal, A., & Saha, S. (2016). A generalized framework for anaphora resolution in Indian languages. Knowledge-Based Systems, 109, 147-159.

[37]. Sakti, S., Paul, M., Finch, A., Sakai, S., Vu, T. T., Kimura, N., & Wutiwiwatchai, C. (2013). A-STAR: Toward translating Asian spoken languages. Computer Speech Language, 27(2), 509-527.

[38]. Nongmeikapam, K., & Bandyopadhyay, S. (2012). A transliteration of crf based manipuri pos tagging. Procedia Technology, 6, 582-589.

[39]. Eryiğit, C., Köse, H., Kelepir, M., & Eryiğit, G. (2016). Building machine-readable knowledge representations for Turkish sign language generation. Knowledge-Based Systems, 108, 179-194.

[40]. Dawa, I., Aishan, W., & Dorjiceren, B. (2014). Design and Analysis of a POS Tag Multilingual Dictionary for Mongolian. IERI Procedia, 7, 102-112.

[41]. Kim, S., Yoon, J., Seo, J., & Park, S. (2012). Improving Korean verb–verb morphological disambiguation using lexical knowledge from unambiguous unlabeled data and selective web counts. Pattern Recognition Letters, 33(1), 62-70.

[42]. Na, S. H., Kim, H., Min, J., & Kim, K. (2018). Improving LSTM CRFs Using Character-based Compositions for Korean Named Entity Recognition. Computer Speech & Language.

[43]. Tomaselli, M. V., & Gatt, A. (2015). Italian tag questions and their conversational functions. Journal of Pragmatics, 84, 54-82.

[44]. Khan, I. A., & Choi, J. T. (2016). Lexicon-corpus Based Korean Unknown Foreign Word Extraction and Updating Using Syllable Identification. Procedia Engineering, 154, 192-198.

[45]. Carneiro, H. C., França, F. M., & Lima, P. M. (2015). Multilingual part-of-speech tagging with weightless neural networks. Neural Networks, 66, 11-21.

[46]. Narayan, R., Chakraverty, S., & Singh, V. P. (2014). Neural network-based parts of speech tagger for Hindi. IFAC Proceedings Volumes, 47(1), 519-524.

[47] Sodhar, I. Naz , Jalbani, A. Hussain Buller, A. Hafeez, (2020). An Empirical and Statistical Study on POS Tagging of Sindhi Social Media text, FOURRAGES 241(1)

[48] Jahara F., Adrita B, MD. Asif Iqbal, Avishek Das, Omar S., Hoque I,M. Moshiul ID , and Sarker ,Iqbal H.(2021) , Towards POS Tagging Methods for Bengali Language: A Comparative Analysis.