ThASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 4. Issue 1 (2024)
https://doi.org/ 10.62019/abbdm.v4i1.111.

Part-Of-Speech Tagging for Balochi Language: A Data driven application of Conditional Random Fields

Sami Ullah, Najma Imtiaz Ali, Shah Murad Chandio, Imtiaz Ali Brohi*, Barkat Ali Laghari

Chronicle

Article history

Received: February 12, 2024
Received in the **revised format:** March 8, 2024
Accepted: March 13, 2024
Available online: March 18, 2024

Sami Ullah, Najma Imtiaz Ali, and Shah Murad Chandio are currently affiliated with the Institute of Mathematics and Computer Science, University of Sindh, Jamshoro, Pakistan.
Email: sami.danish417@gmail.com
Email:najma.channa@usindh.edu.pk
Email: sm.chandio@usindh.edu.pk

Imtiaz Ali Brohi and Barkat Ali Laghari are currently affiliated with Government College University Hyderabad, Pakistan.
Email: brohiimtiaz@hotmail.com
Email:dr.barkatali.laghari@gcuh.edu.pk

***Corresponding Author:**

Abstract

Parts-of-Speech (POS) tagging involves the assignment of the correct part of speech or lexical category to individual words within a sentence in a natural language. This procedure holds significant in the field of Natural Language Processing (NLP) and find utility across a variety of NLP applications. Commonly, it constitutes the initial phase of natural language processing. Subsequent stages may encompass additional tasks such as chunking, parsing and more. Balochi stands as the predominant language in Balochistan,, ranking as the fourth most prevalent language in Pakistan. The field of natural language processing for Balochi is still in its nascent stages. In this research, we introduce an algorithm for Balochi part-of-speech tagging, leveraging machine learning techniques. The core of our approach relies on a Conditional Random Field model as the machine learning component. Careful consideration is given to selecting appropriate features for the CRF, taking into account the linguistic characteristics of Balochi. Balochi is currently considered a resource poor language, and thus, the available manually tagged data consists of only approximately 1500 sentences. The tagset used in this study created for research purpose, consisting of 16 different tags. The learning process incorporates tagged data. The algorithm demonstrates a high accuracy rate of 86.78% when applied to Balochi texts. The training corpus comprises 40000 words, while the test corpus contains 10000 words.

Keywords: Part-of-Speech tagging, Balochi language, Balochi Corpus, Balochi tag set.

INTRODUCTION

Parts of speech tagging involves categorizing the words in a given text based on their definitions and the context of the sentences in which they appear (Anastasopoulos et al., 2010). It is a crucial initial step for numerous Natural Language Processing (NLP) applications. Typically, research in this domain falls into one of three categories: statistical, machine learning, or rule-based approaches (Plank, 2016). Various models employed in the statistical approach include Conditional Random Fields (CRFs), Hidden Markov Models (HMMs), Maximum Entropy Markov Models (HEMMs), among others (Eskander & Collins, 2020). An alternative approach utilizes the rule-based method, which entails devising rules derived from an examination of the language's linguistic aspects. The application of these rules is specifically targeted towards the test corpus (Osenova et al., n.d.). In contrast, tools based on statistical learning typically view the tagging problem as a task of classification. They are not language-specific and may struggle

while it comes to disambiguating words with multiple meaning, as well as handling unknown words that were not present within the training corpora (Dalai. Tusarkanta, n.d.). These tools are dependent on probabilities but do not possess semantic comprehension of the language, which can impact the accuracy of their tag assignments. Additionally, statistical tools require a sizable annotated corpus (Tian et al., 2020). However, they excel at tagging words, with a focus on both familiar and unfamiliar elements, the accuracy by utilizing probabilities of similar tags within specific context and integrating relevant features from the training data (Eskander & Collins, 2020)(Hardie, 2003). Conversely, purely rule-based systems encounter difficulties when encountering unknown words or words that do not adhere to any predefined rules. These systems are unable to predict or suggest likely tags in such cases and may crash when faced with unknown words.

Therefore, to achieve high accuracy using the rule-based approach, an extensive set of rules must be established to account for various scenarios and exceptions (Ding et al., 2018). There is another category of tools known as hybrid systems, which often outperform purely rule-based or statistical approaches. These hybrid tools combine the probabilistic features of statistical tools with language-specific rules applied during post-processing (Ball & Garrette, 2018). An effective strategy involves generalizing language-specific rules and transforming them into features that can be seamlessly integrated into statistical tools (Dalai. Tusarkanta, n.d.). However, a challenge lies in finding the right balance of features and selecting them carefully to ensure accuracy. The greater the number of language specific features that are developed and integrated, the more significant the possibility of attaining enhanced accuracy in the system.

LITERATURE REVIEW

In the past, diverse methods have been utilized for Part-of-speech tagging. Certain scholars have directed their attention towards a rule-based approach inspired by linguistic princples, as exemplified by Brill (Dewi & Santoso, 2020). In the field of machine learning, previous research has primarily relies on two key approaches for sequence labeling. The initial approach involves utilizing a generative probabilistic model of k-order for paired input sequences, such as Hidden Markov Model(HMM) or Multilevel Markov Model (Dewi & Santoso, 2020)(Neale et al., n.d.). The combination of generative and classification models is exemplified in the usage of Conditional Random Fields (CRFs). Much like classification models, CRFs are capable of dealing with multiple statistically correlated input features and are trained in discriminative manner. Additionally, similar to generative models, CRFs are able toc make decisions at various positions within a sequence to achieve an optimal labeling for the entire sequence.

The initial application of CRFs to shallow parsing for English, specifically NP chunking, was demonstrated by Lafferty et al. using the WSJ corpus, resulting in a reported performance of 94.38% (Neale et al., n.d.). For Urdu, CRFs were first employed for POS tagging and chunking by Abdul Wahab Khan et al, yielding reported performances of 84.59% (Rajper et al., 2021)(Javed et al., 2021). furthermore, Lafferty's research showcased that CRFs outperform related classification models and Hidden Markov Models (HMMs) in both synthetic data experiments and POS-tagging tasks (Meftah et al., 2010). Supervised learning approaches have been employed in several POS taggers, utilizing both word instances and tagging rules. These taggers have reported precision levels exceeding 96% for English (Liao et al., 2020). However, in the case of languages like Urdu and other South

Asian languages, there is a scarcity of tagged corpora available (Javed et al., 2021; Khan et al., 2018). Furthermore, the increased morphological complexity of these languages poses a challenge when attempting to attain results that are comparable to those obtained in previous studies conducted on English (Hardie, 2003).

CONDITIONAL RANDOM FIELD

The formulation of Conditional Random Fields (CRFs) by Charles Sutton can be described as follows. Consider a factor graph G over Y, where Y represents the target variable. If, for a given fixed value of x, the distribution p(y | x) can be factorized based on the graph G, then it deemed a Conditional Random Field (CRF). in other words, every conditional distribution p(y | x) can be viewed as a CRF, even if the factor graph is trivial. Let F = {A} represent the collection of factors in G, where each factor conforms to the exponential family format (Patel et al., 2008). In this case the conditional distribution is depicted as follows

$$p(y|x) = \frac{1}{Z(x)} \exp\{\sum_{k=1}^{K(A)} \lambda_{A\,k} f_{A\,k}(y_A, x_A)\}$$

In this context, X represents a probabilistic variable that encompasses unlabeled data sequences, while Y represents a probabilistic variable encompassing sequences with their respective labels. Each component Yi of Y is considered to take values from a limited set of labels Y. For instance, X could represent sentences in a human language, in addition to Y could represent the part-of-speech tags associated with those sentences. The probabilistic variables X along with Y are conjointly distributed. Nevertheless, in the context of a discriminative framework, we create a Conditional model p(Y | X) by utilizing paired observation and label sequences, without the need for explicitly modeling the marginal distribution p(X). CRFs refer to Conditional Random Fields, which are used to assign label sequences to input sequences based on the conditional probability distribution p(Y | X). Lafferty describes the conditional probability of a specific label sequence Y, given an observation sequence X, as the normalized product of potential function (Xiao et al., 2017). Each potential function has the following form:

$$\exp\left(\sum \lambda jtj(Yi-1, Yi, X, i) + \sum \mu ksk\,(Yi, X, i)\right)$$

In the given equation, tj(Yi-1, Yi, X, i) corresponds to a transition feature function, considering both the observation sequence and the labels at positions i and i-1 in the label sequence. Additionally, sk(Yi,X,i) represents a state feature function that takes into account the label at positon I and the observation sequence. The value of parameters λj and μk are determined through the estimation process using the training data (Fu et al., 2022).

$$Fj(Y, X) = \sum fj\,(Yi-1, Yi, X, i)$$

the equation introduces fj(Yi-1,Yi,X,i) as placeholders for state function represents either a state function s(Yi-1,Yi,X,i) or a transition function t(Yi-1,Yi,X,i). This formulation enables us to express the probability of an observation sequence X producing a label sequence Y in the following manner:(Dalai. Tusarkanta, n.d.), (Khan et al., 2018).

$$P(Y|X,\lambda) = \left(\frac{1}{Z(X)}\right) \exp\left(\sum \lambda j\, Fj(Y,X)\right)$$

In the given expression, the normalization factor Z(X) is introduced.

TAGSET

We creat a tagset for Balochi language in this study. The tagset comprises a total of 16 tags and has been specifically tailored to suit the requirements of Balochi language. It encompasses the essential tags needed for part-of-speech tagging, effectively capturing the nuances of language granularity. The tags are organized into five primary groups, with the noun group encompassing general nouns, spatial or temporal nouns, and proper nouns. The verb group comprises both main verbs and auxiliary verbs. Additionally, there is a category for verb and noun modifiers, including adverbs, adjectives, and quantifiers. Lastly, the tagset includes tags for numbers and cardinals.

Our Approach

A machine learning model-based approach proposed within this research study. It leverages supervised techniques to achieve its goals. A conditional Random Field (CRF) is employed to perform statistical tagging on the test corpus. To train the CRF, features are extracted from tagged data. The utilization of well-designed features with the CRF yields significantly higher accuracy compared to other models. The underlying idea is to convert Balochi linguistic rules into features for the CRF, combining the advantages of rule-based and statistical approaches. Nevertheless, because of the constraints on control and flexibility, it is not possible to incorporate all the features into the CRF. Consequently, after the CRF processing, error analysis is conducted. Based on the identified errors, general and language-specific rules are formulated. These rules are then transformed into new features, which are subsequently applied to enhance the CRF's accuracy. Linguistically, Balochi is a language characterized by free word order.(Sanjrani et al., 2020) It exhibits partial agglutination, allowing a maximum of four suffixes to append to the primary root. In Balochi, words are capable of multiple senses, each associated with different tags. For instance, the word 'بال' can function as a particle meaning 'to fly', or as a connective meaning 'to burn\ on the light'. Similarly, 'دور' can serve as a noun denoting 'period', and adjective indicating 'Fashion', or a verb signifying 'to jump'.

In Balochi, it is optional for postpositions to be connected to the head word. For example, one can write either 'من ءَ' or 'منا', both meaning 'من ءَ (to me)'.

Furthermore, this language exhibits the ability to omit words within sentences. For instance:

Sentence: "من ءَ دور کنگی ءٔ چراگ بالگی انت ۔"

Literal: 'من' 'ءَ' 'دور' 'کنگی' 'ءٔ' 'چراگ' 'بالگی' 'انت' '۔'

Tags: 'PRON' 'ADP' 'VERB' 'VERB' 'CONJ' 'NOUN' 'VERB' 'AUX' 'PUNCT'

In this scenario, the verb 'دور' can be omitted, resulting in the next word 'کنگی' assuming the role of the verb.

The features utilized in the CRF model include suffixes, prefixes, and numbers, among others. For example, words ending in the suffix 'اگ', such as 'چراگ', are labeled as NOUN. The Conditional Random Field (CRF) learns by observing tagged words that share the same suffix within the training dataset. The window size for capturing suffix information is set to 4. This approach allows for the exploration of word forming information. Similarly, when words like 'چم' and 'پونز' appear within the training corpora, the CRF observed the associated prefix and applies the appropriate labels to other words along with the same prefix. This methodology facilitates the exploration of stem information. Additionally, if a token represents a number, it is labeled as NUM, and if it contains a number, it is classified as NUMBER.

EXPERIMENTS

It the beginning we developed a rule-based tagging algorithm to process the test data. Our approach integrated machine learning and rule-based techniques to facilitate the tagging process. The initial accuracy achieved was 84.90%. However, upon conducting error analysis, we observed that the limited size of the training corpus resulted in a substantial number of unknown words, distributed across various tags. Consequently, the heuristics employed were not sufficiently effective. Next, employed a CRF tool to process the test data, resulting in an accuracy of 85.88%. in the following error analysis, we found that the selected features did not prove to be adequately effective. In order to tackle this issue, we chose particular features that extended the rule-based approach utilized in the earlier code, while also customizing them to suit the characteristics of the Balochi language. This refinement increased the accuracy to 86.78%. however, further attempts to improve the accuracy by adding additional heuristics proved counterproductive. For instance, converting all NUMBER to NUM and excluding certain tags, such as CC, QW, PRP, when tagging unknown words, led to a decrease in accuracy. We also experimented with tagging words based on the possible tags between the surrounding words, but this approach also resulted in reduced accuracy. Additionally, our attempts to leverage the word forming combination of previous and current words as a heuristic proved unsuccessful.

Table 1.
Part-of-Speech Tagging Outcome and Dataset Volum

Training Data	Test Data	Results in %
40000	10000	86.78%

The limited availability of tagged data for Balochi poses several challenges that can significantly impact the effectiveness of the model and reliability of the results. With only approximately 1500 manually tagged sentences available, the training corpus is considered relatively small for a language specific model. This scarcity could hinder the model's ability to generalize well, especially for less common linguistic pattern or nuances present in Balochi. Scarcity of annotated data result in a substantial number of unknown words during testing. These unknown words might belong to diverse linguistic categories, making it challenging for the model to accurately assign appropriate tags. Consequently, the model's performance on unseen or infrequently occurring words may be compromised. Balochi words, capable of multiple senses and associated with different tags, may introduce ambiguities. The limited annotated data might not provide

sufficient examples to help the model disambiguate effectively, leading to potential inaccuracies in part-of-speech assignments. In Summary, the scarcity of annotated data for Balochi restricts the model's exposure to the language's rich linguistic diversity, leading to potential challenges in handling unknown words, linguistic variations, and ambiguities. These challenges should be carefully considered when interpreting the model's results and making further improvements in future work.

Error Analysis

The data analysis presented earlier shows that errors have been identified across all the tags, primarily attributed to the lack of sufficient training data. Approximately 35% of the words in the corpus were unidentified. The Conditional Random Field (CRF) encountered challenges in accurately tagging certain unknown words due to the inherent flexibility of language. Mistakes occurred when utilizing the features and probabilities associated with the CRF model. For example, a significant number of errors were observed when incorrectly assigning the tag of an adjective to a noun. Here is an illustration:

"جنگل" "NOUN" "چنچو" " ADJ" "زیبا" "NOUN" "انت" "AUX"

Table .
Verification of Errors Occurring in All tags due to Limited Training Data

Actual tag	Assign tag	Counts
VERB	ADV	23
NOUN	PROPN	42
PRON	NOUN	13
PROPN	NOUN	61
ADP	ADJ	2
ADJ	ADV	17
ADV	NOUN	43
CONJ	ADP	32
PREP	ADP	43
NUM	ADJ	12

In the given instance, the term " زیبا" is incorrectly labeled as "NOUN ". Despite being an adjective, the word is labeled as a noun due to its classification as an unfamiliar word. In Balochi language, adjectives have the flexibility to appear both before and after nouns, resulting in the equal likelihood or dependency on the frequency of both "NOUN" or "PROPN" tags for the unknown word in the training corpus. Furthemore, the probability of it being labeled as a noun rises when the following word acts as an adjective. Instances of two consecutive adjectives in the training corpus are rare. Additionally, the presence of a QF before the unknown word increases the likelihood of it being tagged as a noun rather than an adjective. The correction of these errors is only possible if the word is somehow recognizable or known. Another category of errors pertains to the difficulty in performing Named Entity Recognition, which continues to be an unresolved challenge.

By delving into the specific ways identified errors impact the language's overall understanding, this analysis provides a nuanced perspective on the limitations of the

model and guides future efforts towards more accurate and context-aware part-of-speech tagging in Balochi.

CONCLUSION

We conducted training on a Balochi language-specific Conditional Random Field (CRF) model, achieving an accuracy of approximately 86.78%. According to our experiments, we noticed that integrating language-specific rules into CRF features can notably improve the accuracy, potentially resulting in even higher levels of precision. The CRF model is trained using a combination of labeled data consisting of 40000 words. Drawing from the identified errors, it can be inferred that enlarging the training data size reduces the occurrence of unknown words in the test corpus, consequently enhancing accuracy. Additionally, leveraging machine-readable resources such as dictionaries and morphological data, whenever available can further aid in improving accuracy.

FUTURE RECOMENDATIONS

In future, we have plans to manually tag additional training data. Moreover, we aim to develop language resources specifically for Balochi, which will contribute to improving the Tagger's performance. By increasing the volume of training data, we anticipate a substantial improvement in accuracy.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the supervisors and their different affiliations.
Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.
Availability of data and material: In the approach, the data sources for the variables are stated.
Authors' contributions: Each author participated equally to the creation of this work.
Conflicts of Interests: The authors declare no conflict of interest.
Consent to Participate: Yes
Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Anastasopoulos, A., Lekakou, M., Quer, J., Zimianiti, E., Debenedetto, J., Chiang, D., & Fabra, U. P. (2010). *Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource*. *2005*.

Ball, K., & Garrette, D. (2018). *Part-of-Speech Tagging for Code-Switched , Transliterated Texts without Explicit Language Identification*. 3084–3089.

Dalai. Tusarkanta. (n.d.). *Part-of-Speech Tagging of Odia Language Using statistical and Deep Learning-Based Approaches*. *1*(1).

Dewi, N. P., & Santoso, J. (2020). *Combination of Genetic Algorithm and Brill Tagger Algorithm for Part of Speech Tagging Bahasa Madura*. *7*(October), 38–42.

Ding, C., Utiyama, M., & Sumita, E. (2018). *NOVA : A Feasible and Flexible Annotation System for Joint Tokenization and Part-of-Speech Tagging*. *18*(2). https://doi.org/10.1145/3276773

Eskander, R., & Collins, M. (2020). *Unsupervised Cross-Lingual Part-of-Speech Tagging for Truly Low-Resource Scenarios*. 4820–4831.

Fu, W., Shao, P., Dong, T., & Liu, Z. (2022). *Novel Higher-Order Clique Conditional Random Field to Unsupervised Change Detection for Remote Sensing Images*. *Cd*, 1–21.

Hardie, A. (2003). *Developing a tagset for automated part-of-speech tagging in Urdu Andrew Hardie Department of Linguistics and Modern English Language , University of Lancaster*. 1–11.

Javed, T. A., Shahzad, W., & Arshad, U. (2021). Hierarchical Text Classification of Urdu News using Deep Neural Network. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *37*(4).

Khan, W., Daud, A., Nasir, J. A., Amjad, T., Arafat, S., Aljohani, N., & Alotaibi, F. S. (2018). Urdu part of speech tagging using conditional random fields. In *Language Resources and Evaluation* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s10579-018-9439-6

Liao, H., Zhou, Z., Zhao, X., Zhang, L., Mumtaz, S., Jolfaei, A., Ahmed, S. H., & Bashir, A. K. (2020). Learning-Based Context-Aware Resource Allocation for Edge-Computing-Empowered Industrial IoT. *IEEE Internet of Things Journal*, *7*(5), 4260–4277. https://doi.org/10.1109/JIOT.2019.2963371

Meftah, S., Semmar, N., Sadat, F., & Hx, K. A. (2010). *A Neural Network Model for Part-Of-Speech Tagging of Social Media Texts*. 2821–2828.

Neale, S., Donnelly, K., Watkins, G., & Knight, D. (n.d.). *Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh*. 3946–3954.

Osenova, P., Simov, K., Georgiev, G., Zhikov, V., Sh, T., & Nakov, P. (n.d.). *Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian*.

Patel, C., Gali, K., & Technologies, L. (2008). *Part-Of-Speech Tagging for Gujarati Using Conditional Random*. *January*, 117–122.

Plank, B. (2016). *Distant Supervision from Disparate Sources for Low-Resource Part-of-Speech Tagging*.

Rajper, R. A., Rajper, S., Maitlo, A., & Nabi, G. (2021). *Analysis and Comparative Study of POS Tagging Techniques for National (Urdu) Language and other Regional Languages of Pakistan*. *53*(04), 44–53.

Sanjrani, A. A., Naveed, M. S., & Sajid, M. (2020). *Multilingual OCR systems for the regional languages in Balochistan*. *1*, 2157–2167.

Tian, Y., Song, Y., Ao, X., & Xia, F. (2020). *Joint Chinese Word Segmentation and Part-of-speech Tagging via Two-way Attentions of Auto-analyzed Knowledge*. 8286–8296.

Xiao, L., Wang, R., Dai, B., Fang, Y., Liu, D., & Wu, T. (2017). Hybrid conditional random field based camera-LIDAR fusion for road detection. *Information Sciences*, *0*, 1–16. https://doi.org/10.1016/j.ins.2017.04.048