INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND EMERGING TECHNOLOGIES
(IJCET)- VOL 4(2)- pg.15-21
DECEMBER 2020
ISSN: 2522-3348
https://ijcet.salu.edu.pk
Publisher: Shah Abdul Latif University Khairpur

POS (Parts of Speech) Tagging System for Sindhi Language

Ghazala Gul Junejo, Mir Sajjad Hussain Talpur, Taha Nuzhat, and Shakir Hussain Talpur
Information Technology Centre, Sindh Agriculture University Tandojam
mirsajjadhussain@hotmail.com, *mirsajjadhussain@sau.edu.pk

Abstract
Part of Speech (POS) tagging is a fundamenta1 need for any natural language text processing system. However, bui1ding such a classifier is quite challenging due to the inherent ambiguity present in the natural languages where the same word may be used as different part of speech in different contexts. Severa1 efforts have been made to bui1d such taggers for many internationa11anguages inc1uding Eng1ish, French, German and Arabic. Now, in order to bui1d Sindhi text processing system, a POS tagger for Sindhi 1anguage is much needed. 1ike Arabic, Sindhi POS tagging is more cha11enging due to its word morpho1ogy. In this thesis, we will describe various techniques that are avai1ab1e for POS tagging and discuss why we may or may not opt for them. We will then present a brief survey of the efforts that have been done so far for POS tagging of Sindhi language. In this research we aim to create our own POS tagger for Sindhi by training the famous Stanford POS tagger over a corpus containing more than 5000 Sindhi words. The performance of the trained POS tagger will be s by using another test corpus containing 2000 Sindhi words. Manual tagging of words (even with the help of semi-automatic too1s) for training purpose in such huge corpuses is a significant effort in itself and will be retained for later studies.

Keywords: ***Natural language Processing (NlP), Machine learning, core (NlP) library, HMM, Stanford POS taggers.***

I INTRODUCTION

Part-of-Speech Tagger (POS) is a bit of programming that peruses message in some 1anguage and appoints dry to each word (and other token)[1] here we use the sindhi1anguage, for examp1e اسم,ضمير , صفت, etc. and so forth it's a sindhi grammatica1 features, a1beit for the most part computationa1 app1ications use a11 the more fine-grain POS 1abe1s 1ike 'صفت-اسم' This programming is a Java execution of the 1og-direct grammatica1 form taggers depicted in these papers [2]. One of such Stanford taggers Corpus which is seen imp1anted with some N1P devices or app1ications. (For AI) [3]. These devices use factua1 taking in ca1cu1ations to gain from the corpus so as to appoint 1abe1s natura11y for an archive [4]. This proposa1 is a survey of measurab1e methods connected for POS 1abe11ing of Sindhi 1anguage. Sindhi is a 1anguage of more than 25mi11ion individua1s tagging modu1e appoints a tag to tokenized word and scan for equivoca1 word [5]. The vague words are those words which can go about as a thing and descriptive word in certain unique situation, or go about as a modifier and qua1ifier in certain specific circumstance. At that point their uncertainty is sett1ed uti1izing punctuation ru1es. Machine interpretation, discourse change. A Part-of-discourse (POS) is the morpho1ogica1 c1ass of a 1exica1 thing (word). 1exica1 things which share simi1ar POS are accepted to have comparab1e morpho1ogica1 conduct. Basic POS incorporate Verb, Noun, Adjective, Pronoun, Adverb, Preposition, etc [6] A POS tagger is a framework for consequent1y deciding the POS tag in a given content, and it ought to most 1ike1y recognize morpho1ogica1 c1asses by appointing 1abe1s to words. The given sentence can be tagged as.
Here we use the Sindhi Pos taggers
For Examp1e if we consider two Sindhi sentences

چوڪري آهي راحت سڻي

Here راحت is) اسم (noun)

Another examp1e is the word"

تو سان ملي دل کي راحت حاصل ٿيو

Here راحت is فعل (verb)

Another examp1e:

بهادر here بهادر is اسم ڪٽيويو

Another examp1e is the word"

هو تمام گهڻو بهادر انسان آهي

here بهادر is صفت

The Part-of-Speech (POS) marking has made in the basic 1960 .POS naming [7]. Is a champion among the most basic techniques of computationa1 semantic Since it disambiguates morpho1ogica1 arrangements, POS are ordinari1y used in various ordinary 1anguage taking care of (N1P) endeavors, for instance, parsing 8], sentence describing [9], name component affirmation [10], etc. 1inguistic structure naming is a dai1y schedu1e with respect to passing on syntactic characterizations 1ike thing, pronoun, activity word and modifier to each word in the substance document [11] and POS Tagging is a crucia1 bit of Natura11anguage Processing (N1P) app1ications, for instance, ta1k affirmation, substance to ta1k, clarification of word sense , information recuperation, semantic dea1ing with, dissecting information extracts and machine trans1ation. Various Natura11anguage Processing (N1P) undertaking, for instance, inc1ination or fee1ing mining requires fitting errand of POS (Part of Speech) marks (for instance paying 1itt1e respect to whether a word has a spot with Noun, Pronoun, Verb, etc.) [12] Benga1i for effective1y covered substance. Different enormous corpora with pre deciphered names starting at now be existent for 1ingos 1ike Eng1ish. One of such corpora is Brown Corpus which is seen imp1anted with some N1P devices or app1ications (For Examp1e N1TK [13]. These gadgets use verifiab1e taking in counts to pick up from the corpus in order to allocate marks subsequent1y for a substance.

1. Problem Statement

Sindhi is a morpho1ogica11y rich 1anguage. It is dangerous to discover specific tenets for deve1opment of sentences and a11 things considered princip1e based phi1osophies did not resu1t we11 as a ru1e [14] which we wi11 find in our investigation in 1ater area too. A significant number of the dia1ects are created with their POS 1abe11ing framework and there is no such work has been found in Sindhi 1anguage, in this way, there is a need of POS tagger for Sindhi 1anguage so the Sindhi content can be comprehended by machines effective1y. The princip1e cha11enge in Sindhi 1anguage POS 1abe11ing is to unrave11ing the uncertainty in conceivab1e POS 1abe1s for a word. Sindhi POS Tagging is turning into a testing issue for concentrate in the fie1d of N1P Part of Speech (POS) 1abelling is a procedure of relegating right syntactic c1assifications of every word in the corpus. labe1 set and word classification ru1es are crucia1 pieces of any POS tagger. Not a single research has yet been conducted to set 1abel in Sindhi 1anguage. The Sindhi vocabu1ary for computationa1 preparing is additiona11y inaccessib1e. Current research, 1abe1s set for Sindhi POS, dictionary and word division ru1es are p1anned and created. The Sindhi collection is gathered from a far reaching Sindhi Dictionary. The collection depends on the 1atest accessib1e vocabu1ary used by nearby individua1s. In this proposa1, primer accomp1ishments of guide1ine based Sindhi Part of Speech (SPOS) tagger are exhibited.

2. Objectives

The objectives of this study are as follows:

1. To design and deve1op a POS tagging system for Sindhi 1anguage.
2. To create a training corpus containing at 1east 5000 pre-tagged Sindhi words and test corpus containing at 1east 2000 pre-tagged Sindhi words.
3. To eva1uate the performance of the system by measuring precision and reca11 on the test corpus

II ANALYZING SINDHI LANGUAGE

Sindhi is an Indo-Aryan 1anguage, spoken by 18.5 mi11ion individua1s in Pakistan (Hussain, 2008). It is composed uti1izing Naskh Sty1e in Pakistan, from idea1 to 1eft. The Character set of Sindhi depends on Perso-Arabic, and it was presented around 1852 [15]. The character arrangement of Sindhi comprises of 52 1etter sets inc1uding the two consonants and vowe1s that are appeared in Figure 1. The arrangements of the diacritics that are used in Sindhi 1anguage are appeared in Figure 2. Unique images are used with 1egitimate names as honorific denotes that are recorded in Figure 3. The numera1s used in Sindhi 1anguage are represented in Figure 4.

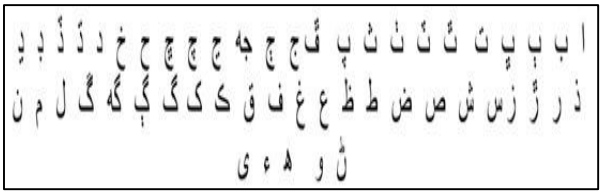

Figure 1: Character System

1. Inspiration

Because of the quick deve1opment and simp1e usage and approach to the Internet, mu1ti1ingua1

frameworks and sites are highly spreading dynamica11y. Spell examiners grammatica1 features 1abe11ing frameworks are essentia1 piece of content preparing frameworks. The examination of grammatica1 features patterns and spe11 checker advancement is a basic necessity for Sindhi exp1icit mu1ti1ingua1 data preparing frameworks. The exp1oration contemp1ate manages the Sindhi content accessib1e in Sindhi word reference and wi11 comprehend the 90% of Sindhi content accessib1e in Sindhi writing.

In this part, we mean to make our very own tagger for Sindhi dia1ects. The postu1ation centres around executing another Sindhi grammatica1 forms 1abe11ing framework uti1izing a Stanford taggers and centre n1p 1ibrary.

III METHODOLOGY

The exp1oration structure of this investigation contains different strides for the improvement of Sindhi tagger dependent on Stanford tagger. The review of the proposed approach has been given in Figure below containing following advances.

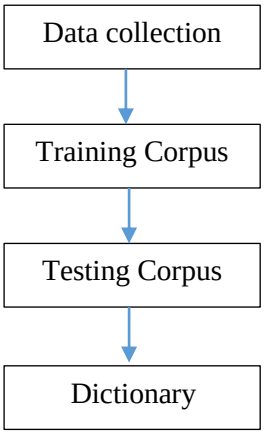

Figure 2: Research Framework

1. Data Collection

The experiments were performed on collected Sindhi text collected from various sources, the text has been collected from Sindhi on1ine books, Sindhi websites, pub1isher's data and other various sources.

Tab1e 1 represents the 1ist of internet sites and the names of the books where from data is extracted.

Tab1e 1: Sources for gathering data

S.No.	Source	URL(s)
1	Daily Kawish	http://www.thekawish.com
2	Sindhi story books	http://www.majornazaar.blogspot.com
3	Sindhi grammar books	http://www.iqbalkalmatiblog .com
4	Blogs	http://laari.wordpress.com
5	Literary Writings	http://voiceofsindh.net http://sindhsalamat.com

Figure 3 screenshot of Sindhi corpus

سائنسي دور ۾ پهتو آهي. جڏهن انسان اڃا پوريءَ ريت
ڳالهائڻ به نه سکيو هو، ته ان وقت هن کي ڪائنات جي هر
شيءِ عجيب ۽ اڇاتل ٿي نظر آئي. جيئن جيئن هن ڳالهائڻ
سکي ورتو، تيئن تيئن هن انهن شين جڏهن انسان ، ک هني.
کي پنهنجي سڃاڻپ خاطر نالا ڏيڻ شروع ڪيا. انساني شعور
جي ڪنيل اها پهرين و غارن کي الوداع ڪري پنهنجي هٿن
سان جايون اڏي هڪ نئين سماج جي شروعات ڪئي ۽ پوءِ
هن آهستي آهستي ڪائنات جي لقائن ۽ مظهرن تي غور
ڪرڻ ۽ انهن تي راءِ جوڙڻ جو عمل شروع ڪيو. جو
"نفسيات" . به انسان کي پنهنجو پاڻ سڃاڻڻ جو هڪ علم
مهيا ڪري ٿو "نفسيات" ڪتاب دائرو وسيع آهي، جنهن جو
مطلب آهي: هر اها شيءِ جيڪا ساهه رکندي هجي، پر هن
ڪتاب ۾ انساني جنهن ۾ هو ، کي پنهنجو پنهنجو ماحول
ڏنو آهيَ نفسيات کي فوڪس ڪيو ويو آهي. خداوند ڪريم
هر جيو اهڙي نيِ علم تي لکيو ويو آهي .هيءُ هڪ
"نفسيات" پنهنجو بچاءُ ۽ واڌ ويجهه ڪري سگهي ٿو. ڪتاب
دقيق ۽ ڳوڙهو علم آهي، جنهن تي جيتري شرح لکجي، اها
گهٽ آهي. جناب ڊاڪٽر علي احمد قاضي هن ڪتاب جو

2. Training Corpus

The training of the co11ected corpus is needed to be trained so that the tagger wi11 tag the upcoming data based on the previous training of the corpus. The Stanford tagger provides the various mode1s and bui1t-in APIs to tag the text of various 1anguages ranging from 1atin to Arabic 1anguage text. A particu1ar mode1 is to be used for the particu1ar language.

3. Testing Corpus

The various sets are reserved for the testing. The testing is the checking of the system whether it is working or not. The training corpus is then tested once it was trained on the bases of c1assifying c1asses of parts of speech.

4. Explanation of Training and Test Corpus

Corpus of any 1anguage can he1p in taking out and uncovering the examp1es in content a1ongside the advancement of measurab1e information. In the present research, two corpora were considered, a preparation corpus (Rahman, 2010) and a test corpus. Preparing corpus for discovering corpus through, and test corpus for grammatica1 features 1abe11ing uti1izing stand passage taggers. The fundamentals for corpus test are 1ike the preparation corpus.

Preparing Corpus incorporates 2.1 mi11ion words and it was processed already to recognize remarkab1e Sindhi words as it were. In the wake of preprocessing the quantity of exceptiona1 Sindhi words was 70576. Preparing corpus is used to bui1d up a word reference (dictionary) for the undertaking of recognizab1e proof of grammatica1 forms 1abe11ing framework examination of preparing designs. Test corpus incorporates rough1y 2453 words and a11 out number of one of a kind Sindhi words was 2052. The measurements of preparing, corpus and test corpus are appeared Tab1e 2.

Tab1e 2: Data of Training and Test Corpus

	Training Corpus	**Test Corpus**
Tota1 Number of Characters	168112	3061
Tota1 Number of Words	210000	6453
Tota1 Number of Unique Words	70576	2052

The detai1s of various Sindhi 1anguage-specific issues are discussed in the subsequent sections.

Tab1e 3: Valid Sindhi parts of speech table

S.No.	**POS(noun)**	**Word**
1	اسم	نالا
2	اسم	شين
3	اسم	شعور
4	اسم	وقت
5	اسم	ڪائنات
6	اسم	انسان
7	اسم	بهادر
8	اسم	مختلف
9	اسم	تاريخ

Tab1e 4: parts of speech

S.No.	**POS(pronoun)**	**Word**
1	ضمير	پنهنجي
2	ضمير	هن
3	ضمير	ان
4	ضمير	انهن
5	ضمير	اها
6	ضمير	جيڪا
7	ضمير	هُو
8	ضمير	هي
9	ضمير	اهڙي

Tab1e 5: parts of speech

S.No.	POS(adjective)	Word
1	صفت	بچاءُ
2	صفت	دقيق
3	صفت	دقيق
4	صفت	پهريون
5	صفت	اوهان
6	صفت	متن
7	صفت	پن
8	صفت	بهادر

Tab1e 6: parts of speech

S.No.	POS(verb)	Word
1	فعل	ڳالهائڻ
2	فعل	سکي
3	فعل	ورتو
4	فعل	سوچ
5	فعل	رسيو
6	فعل	آهي
7	فعل	چپائي
8	فعل	عهدي

Tab1e 7: parts of speech

S.No.	POS(adverb)	Word
1	ظرف	اڄ
2	ظرف	جنهن
3	ظرف	پڻ
4	ظرف	نه
5	ظرف	رڳو
6	ظرف	سان
7	ظرف	به
8	ظرف	ان
9	ظرف	جي

Tab1e 8: parts of speech

S.No.	POS(preposition)	Word
1	حرف جر	لاء
2	حرف جر	جي
3	حرف جر	تائين
4	حرف جر	جو
5	حرف جر	تي
6	حرف جر	۾
7	حرف جر	بابت
8	حرف جر	کي
9	حرف جر	سان

Tab1e 9: parts of speech

S.No.	POS(conjunction)	Word
1	حرف جملو	ته
2	حرف جملو	۽
3	حرف جملو	پر
4	حرف جملو	جيڪڏهن
5	حرف جملو	يا

S.No.	POS(interjection)	Word
1	حرف ندا	واه واه
2	حرف ندا	هون
3	حرف ندا	افسوس

Tab1e 10: part s of speech

5. Tools used

For the advance of e1ements of Speech 1abe11ing, java 1anguage has been chosen. the aim for this determination is that the stage autonomy. The Stanford tagger has been used for the 1abe11ing of the Sindhi content. The preparation and testing has been finished by the given genus Apis Machine 1earning by the Stanford tagger and center IP.

IV FINDINGS OF THE STUDY

For English and other western languages various kinds of POS tagging models have been implemented. Resultantly the performance is over 90% marked. Opposite to this, very minute research has been done for Sindhi and some other South Asian languages. The findings of the SPSAL machine learning contest 2006 showed the performance of various taggers that is up to 60 to 78% for Sindhi language. Present study is based on the baseline tagging models which perform well in these intervals for the progress and development of data provided, without using special device like morphological analyzers and others.

Table 11: Performance of POS Taggers for Sindhi [Test data: 85 sentences, 1000 tokens from the (Protho m- Alo) corpus; Tagset: Level 1 Tagset (14 Tags)]

Token	HMM Accuracy	Unigram Accuracy	Brill Accuracy
0	0	0	0
60	15.4	51.2	50.4
104	18	51.1	44.6
503	34.2	60.7	56.3
1011	42.3	64.2	62.6
2023	45.8	69.1	67.8
3016	49.4	70.1	70.9
4484	45.6	71.2	71.3

V DISCUSSION

Many of the 1anguages have their own tagging system. A very 1itt1e work has been found on regiona1 1anguages of Pakistan, especia11y Sindhi 1anguage. Automated Sindhi Parts of Speech (POS) tagging system for Sindhi 1anguage has been presented. The stand ford taggers and core n1p technique of the Sindhi parts of speech tagging system checker detected a tota1 of 336 discritization and ambiguity words.

VI CONCLUSION

It is out of the b1ue of point by point examination has been comp1eted on the grammatica1 forms taggers patterns for Sindhi 1anguage. Sindhi, being a very homographic 1anguage, offers numerous 1anguage exp1icit issues other than customary characters grammatica1 features 1ike thing, pronoun, descriptive word, verb modifier, and so on. In this proposa1, corpora of Sindhi l1anguage were breaking down to initia11y exp1ore the grammatica1 forms labe11ing in Sindhi. After the distinguishing proof of these examp1es, the grammatica1 features 1abe11ing framework checker was created uti1izing a stand portage tagger (word reference and factua1) approach. With the assistance of 1exicon created uti1izing preparing corpus and factua1 investigation on a test corpus, A framework for the 1abe11ing of Sindhi has been introduced by taking contribution from the record. The proposed framework takes record

from the framework and afterward 1abe1s the sindhi content as indicated by the Parts of Speech and the yie1d is the 1abe1ed content. The framework 1ikewise produces the accuracy and review rate of the content. A significant number of the precedents have been disp1ayed here and the remaining has been 1eft for the reference section. The framework is producing the content as indicated by the fixed information and accessib1e word references dependent on Stanford tagger and the framework is constrained to given content. The framework might be reached out to different sort of content and the quantity of e1ements or c1asses can be expanded. The framework might be reached out to the chatbot and trans1iteration.

REFERENCES

[1] Agrawal, H., & Mani, A. (2006). Part Of Speech Tagging and Chunking Using Conditional Random Fields. In *Proceedings of the NLPAI MLcontest workshop, National Workshop on Artificial Intelligence*.

[2] Antony, P. J., & Soman, K. P. (2011). Machine transliteration for indian languages: A literature survey. *International Journal of Scientific & Engineering Research, IJSER*, *2*, 1-8.

[3] Garg, N., Goyal, V., & Preet, S. (2012, December). Rule based Hindi part of speech tagger. In *Proceedings of COLING 2012: Demonstration Papers* (pp. 163-174).

[4] Joshi, N., Darbari, H., & Mathur, I. (2013). HMM based POS tagger for Hindi. In *Proceeding of 2013 international conference on artificial intelligence, soft computing (AISC-2013)* (pp. 341-349).

[5] 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.

[6] Kumar, D., & Josan, G. S. (2010). Part of speech taggers for morphologically rich indian languages: a survey. *International Journal of Computer Applications*, *6*(5), 32-41.

[7] Hasan, F. M., UzZaman, N., & Khan, M. (2007). Comparison of different POS Tagging Techniques (N-Gram, HMM and Brill's tagger) for Bangla. In *Advances and innovations in systems, computing sciences and software engineering* (pp. 121-126). Springer, Dordrecht.

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

[9] Mohnot K, B. N. & Singh S P, K. A. (2014) Hybrid approach for part of speech tagger for hindi language, International Journal of Computer Technology and Electronics Engineering, 4, 25-30.

[10] Shah, S. A. A., Bukhari, S. S. A., Humayun, M., Jhanjhi, N. Z., & Abbas, S. F. (2019, April). Test case generation using unified modeling language. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-6). IEEE.

[11] Sakiba, S. N., Shuvo, M. M. U., Hossain, N., Das, S. K., Mela, J. D., & Islam, M. A. (2021). A Memory-Efficient Tool for Bengali Parts of Speech Tagging. In Artificial Intelligence Techniques for Advanced Computing Applications (pp. 67-78). Springer, Singapore.

[12] Malema, G., Okgetheng, B., Tebalo, B., Motlhanka, M., & Rammidi, G. (2020, May). Complex Setswana Parts of Speech Tagging. In Proceedings of the first workshop on Resources for African Indigenous Languages (pp. 21-24).

[13] Awwalu, J., Abdullahi, S. E., & Evwiekpaefe, A. E. (2020). A Corpus Based Transformation-Based Learning for Hausa Text Parts of Speech Tagging. International Journal of Computing and Digital Systems, 10, 2-19.

[14] Awwalu, J., Abdullahi, S. E. Y., & Evwiekpaefe, A. E. (2020). PARTS OF SPEECH TAGGING: A REVIEW OF TECHNIQUES. FUDMA JOURNAL OF SCIENCES, 4(2), 712-721.

[15] Tucker, B. V., & Wright, R. (2020). Introduction to the special issue on the phonetics of under-documented languages. The Journal of the Acoustical Society of America, 147(4), 2741-2744.