Reusing Stanford POS Tagger for Tagging Urdu Sentences

Adnan Naseem[1], Muazzama Anwar[1], Salman Ahmed[2], Avais Jan[1], Ahmad Kamran Malik[1]
[1]Deptment of computer science
CIIT Islamabad, Pakistan
[2]Deptment of computer science and software engineering
IIUI Islamabad, Pakistan
anasim.ciit@gmail.com, muazzama.anwar@yhoo.com, contectsani@gmail.com, Jan.avais@yahoo.com, ahmad.kamran@comsats.edu.pk

***Abstract*—Several Natural Language Processing applications in a particular language consider POS tagging a necessary component. To develop a new language specific POS tagger targeting such particular language is a tedious job for unstructured data due to the variation in text, type and complexity of text. For that reason, it impacts the precision of tagging as a result of the variety of a certain language. Current research focused on the thought of reusability of a popular language specific Part of speech tagger, for example, Stanford Part of speech Tagger can be employed for tagging non-Engish phrases. For generalizeability, any translator can be used to translate the sentences, however, a well-known translator, named "Google translator" is considered for sentence translation purpose across the languages. For evaluation perspective, Urdu tweets of a hot political issue "Panama leaks" from twitter.com are extracted. To measure the accuracy, the kappa statistic along with confusion matrix is deliberated. The precision of tagging the Urdu sentences by reusing Stanford Part of speech tagger is 96.05 percent. The respected approach can be globally applied to tagging the sentences of several different languages.**

Index Terms—"Stanford-Part-of-speech Tagger; Google-Translator; Multi-lingual labling";

I. INTRODUCTION

Natural language processing (NLP) is an area focused on the interactions between computer system and humans (natural) languages. Part of speech (POS) tagging is a significant action to a number of NLP tasks considering speech recognition, machine translation, information retrieval, grammar checking, etc. It concerned with reading scripts in one language and assigning grammatical tags (NN, VB, ADJ, and JJ) to every single term in the sentence. It is certainly a fundamental type of syntactic evaluation of a language, that includes numerous implications in NLP. Many part of speech taggers tend to be trained from ("treebanks") for instance, Penn Treebank. Conversely, Stanford part of speech Tagger is the hottest option for the researchers due to its several packages support[1]. For example, GATE, C#/F#/.Net, Docker, Go, Javascript (node.js), Matlab, XML-RPC, PHP, Ruby and Python. As a result of discontinuation of tagging from domains information, as well as nature of Twitter discussions, absence of conventional orthography, and also 140-character size restriction for each and every message ("Tweet") are the Obstacles still experienced.

The growing popularity of social media and user- created web content is producing enormous quantities of text in electronic form. As English is an international language for communication, an abundant source of data is increasing with quick rate comprising some useful information. Because of its massive stuff, Yet, it is a challenging task to filter the helpful information[2]. Little literature has been considered in part of speech tagging as compared to English. To improve the accuracy of tagging many techniques have been explored. These techniques vary from being purely rule based in their approach to being completely stochastic. Stanford POS Tagger is a bit of which checks out the writing in a certain language and assigns tags to every single term. The information besides in English language is also essential. In order to increase the quality in non-English text, English Part of speech taggers have been used again for non English sentences tagging.

Current research comprises of broad literature review of English part of speech taggers was done which are written explicitly for English, where, the Stanford part of speech tagger are reused to tag Urdu sentences. To extract Urdu sentences (tweets) on particular topics, twitter API is used. For additional processing, unbiased randomly selected Urdu tweets are arbitrarily considered, followed by the improvement process. By using "Google translator" such un-biased sentences were translated and retranslated within native and other languages. Latest English part-of-speech taggers were takeout and mentioned in current practice. Though, their results desired to be mentioned in the extensive work of this paper. In order to get tagged-English sentences, these English tweets are introduced to the Stanford part-of-speech tagger. By using Google translator such labled English tweets are re-translated to the original language. The annotations were performed by two human annotators as standard labled tweets. "Kappa statistic with confusion matrix" is used for precision perspectives.

The very next section discusses of extensive literature review. Section III comprises the research methodology. Results are deliberated in IV Section. The final section comprises of Conclusions.

II. BACKGROUND KNOWLEDGE

This section focuses a detailed literature review. The present study varies from existing in terms of generalizability and reusability of part of speech tagger.

A CLE Urdu Parts of Speech [4] is used which considered the basic CLE technique named Urdu Digest Tagged Corpus have accurate result up to 96.8%. Anwar et al [5] used a model named n-gram Markov Model and proposed a "n gram based POS tagger", for Urdu language, where the precision is up to 95%. Jawaid and Bojar [6] proposed Tagger voter for Urdu in which the authors improved parts of speech tagging for Urdu by using "Humayoun's morphological analyzer", "SVM Tool tagger", the tool give accuracy up to 87.98%. Anwar et al[7] tries to solve the aforementioned problem by using the Hidden Markov Model. Sajjad and Schmid [8] done a comparison between four Urdu taggers named "TnT tagger", "TreeTagger", "RF tagger" and "SVM tool". Authors found highest 95.66% accuracy by SVM tool. Hardie[9] propsed first computational part of speech tagset for Urdu, creating one of the necessary resources for the development of a POS tagging system for Urdu.

A rule based Urdu tagging was considered by Hardie [10] by using Unitag architecture. Chatterji [11] proposed multi-lingual NER systems by using "Language specific rules and Maximum Entropy" technique. Nazr et al[12] proposed a novice Urdu part-of-speech tagset by using the Penn Treebank which give accuracy of 96.8%. Ahmed et al [13] Proposed "Named Entity Recognition (NER) system" for Urdu language by using Urdu NER system. Riaz [14] proposed Named Entity Recognition by using technique rule-based Urdu NER algorithm. Singh et al[15] identifies the problem of NER in the context of Urdu by using technique IJCNLP-08 and Izaafats on base of their finding twelve NE are proposed.

Malik and Sarwar [16] Proposed NER on "Conditional Random Field (CRF)" by using the accuracy measures. Adeeba and Hussain [17] developed an Urdu WordNet. Hussain [18] worked on Urdu text to speech. Ahmed and Hautli [19] created a vocabulary based understanding means for Urdu by using "Hindi WordNet" by utilizing transliterators where he locates computational semantics on the basis of the Urdu ParGram sentence structure. Hussain and Afzal[20] run Urdu Computing Standards: Urdu Zabta Takhti UZT 1.01 standard using Unicode as a typical. Riaz et al [21] proposed a "vowel insertion grammar" for Urdu using creating speech synthesis for Urdu language. Khanam et al[22] proposed an automated Part-of-speech tagging simply by using "Entropy (ME) modelling system" , "Morphological analyser(MA)" and stemmer suggested various models "ME", "ME+Suf", "ME+MA", "ME+Suf+MA".

Jawaid et al[23] launch a significant "mono-lingual Urdu corpus" instantaneously labled with POS tags. Ali et al[24] examining the governmental "News dictionary" for ruling Significant Objects, Saliences in the Urdu language making use of "Heuristic based Salience research of Urdu Information dictionary" which give reliability to 85.5percent. Khanam et al [25] work on efficient ways of computational linguistics. With an evaluation in "TnT tagger", "optimal Entropy tagger" and "CRF (Conditional Random Field)". Mukund et al[26] works on Urdu-to-English transliteration using Bootstrap and discover precision around 84.1percent. Munir et al[27] works on analysis of "URDU.KON-TB" in dependency parsing domain. "MaltParser", the procedure familiar with exercise, and assessment information is "Nivre arc-agear algorithm". The investigations reveal "URDU.KON-TB Treebank" is certainly misfit for the dependency analyzing. Siddiq et al[28] suggested "statistical model" utilized in one's research "HMM" alongside "IOB" amount manual comment making use of TnT Tagger in which he discover reliability to 97.52%. Ali et al work[29] on "noun phrase chunker for Urdu" that can be predicated on a "statistical method" by making use of HMM based method which give precision as much as 97.61.

III. METHODOLOGY

The current part of the paper presents the methodology of the study. "Twitter" comprises of enormous data facilitates users with its APIs to extract the useful information. We have considered this process too. With the help of twitter APIs, data exacted for a famous topic "PANAMA CASE".

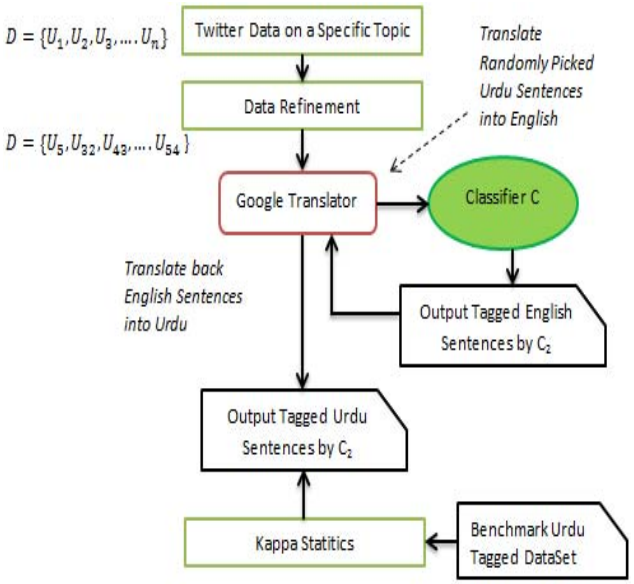

Fig. 1. Research Methodology

Raw information was refined and unbiased 10 Urdu sentences were considered. The randomly picked Urdu sentences were translated into Classifiers native language, i.e., English. Where, the classifier named, Stanford part of speech tagger classifies the English sentences into fifteen unique tags. The output labled English phrases are re-translated to their inventive language that is Urdu in our context. 'Google translator' was considered for the said task as shown in Fig.1. In order to find out the accuracy and precision of the output results, two annotators were considered for synthesizing the benchmark tags. Furthermore, the tags assigned by our system will be compared with respect to the benchmark dataset. For mitigating the factor of biasness, the whole process is repeated three times variations in sample data for more accurate and concrete results.

A very famous social network platform named 'Twitter'[1] where a lot of people communicate with each other via short texts (typically a length of maximum up to 140 characters) named 'tweets'. They conversed a huge number of tweets per day on various topics, especially the political one. We have considered a political topic too for our analysis using keywords

[1]http://twitter.com/

('Panama and PMLN'). Yet, the identical and resembled 'tweets' were filtered out, whereas, the whole tweets were also reviewed by Twitter API[2]. API was being applied by several checks in order to avoid the repetition of tweets. The 'Hash functions' were considered to make sure the unique tweets. Although, the "URLs, twitter connector (@username) and hashtags (#PTI, #PMLN)" were filtered out in the initial stage from tweets and further were considered as a key in 'HashMap'. The 'HASHMAP' keys were mapped to unique tweets. This filtration process resulted 40 percent removed tweets. The remaining tweets were surely considered the unique tweets. However, each tweet was considered to be a unique sentence.

Table 1. Unbiased Randomly Selected Urdu Sentences

S. No	Unbiased Randomly Selected Urdu Sentences
.(1	نوازشریف اور عوام کے دل ایک ساتھ دھڑکتے ہیں۔
.(2	عوام تو بیچاری معصوم ھے اس لیے ھم عدالت عظمی کا فیصلہ ھی مانیں گے۔
.(3	محبت کا یہ عالم ھے کہ سارا پیسہ اور کاروبار ملک سے باہر ھے۔
.(4	یہ آج جہلم نہیں بلکہ جے آئی ٹی کا چہلم تھا۔
.(5	نواز شریف کا باوفا ساتھی مشکل وقت میں بے وفا کیوں؟احتساب کا قانون سب کے لیے یکساں ہونا چاہیئے۔
.(6	جہلم والوں نے پرجوش استقبال کر کے اپنےقائد سے محبت کا حق ادا کردیا۔
.(7	اور تمہارے پاس پاسپورٹ پاکستان کا ہے لیکن جائدادیں برطانیہ میں ہیں۔
.(8	یہ ریلی واہگہ بارڈر کراس کر لے گی۔
.(9	راولپنڈی کے رہائشی شفاقت مرزا کا اپنے قائد سے محبت کا اظہار۔
.(10	کمر میں تکلیف کے باعث ریلی میں شرکت نہیں کی۔
.(11	اگر موت سے ڈرتے ہو تو عوامی ریلیاں مت نکالو۔

A process of random selection of Urdu sentences (Table 1) were considered for generalizability, whereas, the biasness wa mitigate via repeating the process three times. There were numerous English part of speech taggers in the literature, yet, the Stanford part of speech tagger is selected for its generalizability, multi-lingual computer languages support and a wide range of helping audience. Another study focused the extensive literature for the English part of speech taggers. The results shows that Stanford part of speech tagger outclassed the rest of the elected part of speech taggers for accuracy. Selected Urdu sentences were translated into English sentences by considering a multi-lingual translator i.e, 'Google Translator'[3].

The unbiased randomly selected Urdu sentences were being tagged from Stanford part of speech tagger after translation into English by google translator (Table 2).

Table 2. Test Tweets

S.No	Tagged 11 English translated tweets
.(1	Nawaz/NNP Sharif/NNP and/CC the/DT people/NNS 's/POS heart/NN have/VBP shattered/VBN together/RB
.(2	People/NNS are/VBP poor/JJ and/CC poor/JJ ,/, so/IN we/PRP will/MD decide/VB the/DT court/NN order/NN/The/DT love/NN of/IN the/DT love/NN is/VBZ that/IN all/PDT the/DT money/NN and/CC business/NN is/VBZ out/IN of/IN the/DT country/NN
.(3	It/PRP was/VBD not/RB Jhelm/NNP but/CC JIT/NNP 's/POS favorite/JJ today/NN
.(4	Nawaz/NNP Sharif/NNP 's/POS companion/NN ,/, why/WRB is/VBZ he/PRP unhealthy/VBN in/IN difficult/JJ times/NNS ?/.
.(5	The/DT law/NN of/IN accountability/NN should/MD be/VB the/DT same/JJ for/IN everyone/NN ./.
.(6	People/NNS of/IN Jhelum/NNP paid/VBD favor/NN to/TO their/PRP$ leaders/NNS by/IN welcoming/VBG passion/NN
.(7	And/CC you/PRP have/VBP a/DT passport/NN in/IN Pakistan/NNP ,/, but/CC the/DT property/NN is/VBZ in/IN the/DT UK/NNP
.(8	This/DT rally/NN will/MD cross/VB the/DT Wagah/NNP border/NN
.(9	Rawalpindi/NNP resident/NN Shafti/NNP Mirza/NNP expresses/VBZ his/PRP$ love/NN for/IN the/DT leader/NN
.(10	Due/JJ to/TO difficulty/NN in/IN the/DT waist/NN ,/, the/DT rally/NN did/VBD not/RB attend/VB
.(11	If/IN you/PRP are/VBP afraid/JJ of/IN death/NN ,/, do/VBP not/RB leave/VB public/JJ rallies/NNS

A famous translator, i.e. 'Google translator' is considered for translation of sentences across the languages. Urdu sentences were first translated into English sentences, after tagging in native languages by Stanford the part of the speech tagger, such lablelled tweets are re-translated to Urdu as shown in Table 3.

Table 3. Tagged 11 Urdu re-translated Tweets

S.No	Tagged 11 Urdu re-translated Tweets

[2]http://twitter4j.org/en/index.html

[3]https://translate.google.com/

1).	VBP|ہیں VBN|دھڑکتے RB|ایک ساتھ NN|دل POS|کے NNS|عوام CC|اور NNP|شریف NNP|نواز
2).	MD|گے VB|ھی مانیں NN|کا فیصلہ NN|عدالت عظمی PRP|ھم IN|اس لیے VBP|معصوم ھے JJ|تو بیچاری NNS|عوام
3).	VBZ|ھے IN|سے باہر NN|ملک NN|کاروبار CC|اور NN|پیسہ PDT|سارا IN|کہ VBZ|یہ عالم ھے IN|کا NN|محبت
4).	VBD|تھا JJ|چہلم POS|کا NNP|جے آئی ٹی CC|بلکہ RB|نہیں NNPجہلم NN|آج PRP|یہ
5).	IN|کے لیے NN|سب NN|قانون IN|کا NN|؟احتسابWRB|کیوں VBN|بے وفا IN|میں NNS|وقت JJ، مشکلNN|باوفا ساتھی POS|کا NNP|شریف NNP|نواز MD|چاہیئے VB|ہونا JJ|یکساں
6).	VBD|ادا کردیا NN|محبت کا حق TO|سے NNS|قائد $PRP|اپنے IN|کر کے VBG|استقبال NN|پرجوش NNP|والوں نے NNP|جہلم
7).	VBZ|ہیں IN|میں NNP|برطانیہ NN|جائدادیں CC|کا ہے لیکن NNP|پاکستان NN|پاسپورٹ VBP|پاس PRP|تمھارے CC|اور
8).	MD|کر لے گی VB|کراس NN|بارڈر NNP|وابگہ NN|ریلی DT|یہ
9).	VBZ|کا اظہار NN|سے محبت NN|کا اپنے قائد NNP|مرزا NNP|شفاقت NN|کے رہائشی NNP|راولپنڈی
10).	VBD|کی RB|نہیں VB|میں شرکت NN|ریلی JJ|کے باعث NN|تکلیف IN|میں NN|کمر
11).	VB|نکالو RB|مت NNS|ریلیاں JJ|تو عوامی JJ|سے ڈرتے ہو NN|موت IN|اگر

IV. RESULTS

To check out the precision for the subjected part of speech tagger "Kappa Statistic" was used. By hand annotations were considered by using two annotators. A total of fifteen special tags were considered for measurement along with the confusion matrix of "actual tagging" across "predicted tagging". Whereas, the "actual tagging" was performed by the aforesaid 2 annotators.

Table 4. Results

Tags	Actual Results		Predicted Results		Total	Accuracy	Predicted Accuracy	Kappa Statistic	Avg. Accuracy
			Not NN	NN					
		Not NN	103	0	133	0.9699248	0.667137769	0.909646739	
NN	**Actual**	NN	4	26					0.960531128
		Not NNP	120	0		1	0.823619198	1	
NNP	**Actual**	NNP	0	13					
VB		Not VB	128	0		0.9924812	0.934592119	0.885047537	
	Actual	VB	1	4					
VBN		Not VBN	131	0		1	0.97037707	1	
	Actual	VBN	0	2					
VBD		Not VBD	129	0		1	0.941658658	1	
	Actual	VBD	0	4					
MD		Not MD	131	0		1	0.97037707	1	
	Actual	MD	0	2					
VBG		Not VBG	132	0		1	0.985075471	1	
	Actual	VBG	0	1					
POS		Not POS	130	0		1	0.9559048	1	
	Actual	POS	0	3					
NNS		Not NNS	127	0		1	0.913844762	1	
	Actual	NNS	0	6					
RB		Not RB	129	0		1	0.941658658	1	
	Actual	RB	0	4					
IN		Not IN	117	0		1	0.788343038	1	
	Actual	IN	0	16					
VBZ		Not VBZ	128	0		1	0.927638645	1	
	Actual	VBZ	0	5					
VBP		Not VBP	128	0		1	0.927638645	1	
	Actual	VBP	0	5					
JJ		Not JJ	125	4		0.9699248	0.913392504	0.652741514	
	Actual	JJ	0	4					

However, the "predicted tagging" was performed by the proposed system. Each tag was considered for kappa measurement where two accuracies were considered, i.e, random and total accuracy. The following equations (1), (2) and (3) shows the formula used by kappa statistic.

$$Kappa = \frac{Total\ Accuracy - Random\ Accuracy}{1 - Random\ Accuracy} \quad (1)$$

Whereas,

$$Total\ Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + \ False\ positive + False\ negative} \quad (2)$$

$$Total\ Accuracy = \frac{(True\ Negative + False\ Positive) * (True\ Negative + False\ negative) * (True\ positive + False\ negative) * (True\ positive + False\ Positive)}{Total * Total} \quad (3)$$

Additionally, after the addition of all the extracted results the average was considered. Reusing of Stanford Part of speech tagger tags the Urdu sentences with 96.05 percent accurate predictions as shown in Table 4. In order to make the process of selection of Urdu sentences unbiased, random sentences were taken thrice and average were considered. The results are more accurate than the ordinary domain specific Urdu part of speech taggers.

V. CONCLUSIONS

Part of speech tagging is a necessary component of natural processing languages. Due to the less academic focus and diversity of a particular language, it is very hard to develop a domain specific high precise and more accurate part of speech tagger. Therefore, the concept of reusing Stanford part of speech tagger is proposed to tag multi-lingual sentences. For generalizeability, any translator can be used to translate the sentences, however, a well-known translator, named "Google translater" is considered for sentence translation purpose across the languages. For evaluation perspective, Urdu tweets from a political issue "Panama leaks" from twitter.com were extracted. To measure the accuracy, 'the kappa statistic along with confusion matrix' is deliberated. The precision of tagging the Urdu sentences by reusing Stanford Part of speech tagger is 96.05 percent. The respected approach can be globally applied to tagging the sentences of several different languages.

Similarly to other studies, this study has also some restrains. In translation and re-translation of native to non-native and non-native to native language, many translators come up with different translations of the same sentences. Furthermore, when the same translator re-translates the origional text gets drastic outomes. Current work focused on translation has done using mapping the words. For example, this is my Book (ye mara ketab ha) (this,ye), (is,ha), (my,mara) and (book, ketab). The whole process could be made easy by customizing the translator for particular language. Random selection of sentences is another constrain of this work. Though sample sentences were taken three times but still the results were almost the same.

In the current study, short text was used, while text which is not from the twitter will be used in the future paper. A comparison of extensive English POS taggers planned to carried out for prioritized top most Urdu labelled tagger. Additionally, for validation purposes, sample data which is other the twitter will be considered. The existing procedure might be extended to use a tagger for several language tagging in order to get useful information. So, for different languages a generic procedure will be considered in the upcoming work. Furthermore, all languages has several diverse level, that's why, the identical procedure can be considered for to many languages in order to abstain from the construction of a new difficult taggers.

REFERENCES

[1] B. Rehman, Z. Halim, and M. Ahmad, "ASCII Based GUI System for Arabic Scripted Languages: A Case of Urdu," *Int. Arab J. Inf. Technol.*, vol. 11, no. 4, 2014.

[2] M. J. O. C. S. (MJCS), "Artificial Neural Network-Based Speech Recognition Using Dwt Analysis Applied On Isolated Words From Oriental Languages." .

[3] "CLE Store." [Online]. Available: http://www.cle.org.pk/clestore/postagger.htm. [Accessed: 26-Aug-2017].

[4] Naseem, A. et al."Tagging Urdu Sentences from English POS Taggers" *International Journal Of Advanced Computer Science And Applications*, Vol. *8*(10),2017, 231-238.

[5] Anwar, W., Wang, X., Li, L., & Wang, X. L. (2007, August). A statistical based part of speech tagger for Urdu language. In *Machine Learning and Cybernetics, 2007 International Conference on* (Vol. 6, pp. 3418-3424). IEEE.

[6] Bojar, B.J.O., 2012, December. Tagger Voting for Urdu. In *24th International Conference on Computational Linguistics* (p. 135).

[7] W. Anwart et al. "Hidden markov model based part of speech tagger for urdu.," 2015.

[8] Sajjad, H. and Schmid, H., 2009, March. Tagging Urdu text with parts of speech: A tagger comparison. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 692-700). Association for Computational Linguistics.

[9] Hardie, A., 2003. Developing a tagset for automated part-of-speech tagging in Urdu. In *Corpus Linguistics 2003*.

[10] Hardie, A., 2004. *The computational analysis of morphosyntactic categories in Urdu* (Doctoral dissertation, Lancaster University).

[11] Saha, S.K., Chatterji, S., Dandapat, S., Sarkar, S. and Mitra, P., 2008, January. A hybrid approach for named entity recognition in indian languages. In *Proceedings of the IJCNLP-08 Workshop on NER for South and South East Asian languages* (pp. 17-24).

[12] T. Ahmed *et al.*, "The CLE Urdu POS Tagset."

[13] Naz, S., Umar, A.I., Shirazi, S.H., Khan, S.A., Ahmed, I. and Khan, A.A., 2014. Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language. *Research Journal of Applied Sciences, Engineering and Technology*, *8*(10), pp.1272-1278.

[14] Riaz, K., 2010, July. Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 named entities workshop*(pp. 126-135). Association for Computational Linguistics.

[15] Singh, U., Goyal, V. and Lehal, G.S., 2012. Named Entity Recognition System for Urdu. In *COLING* (pp. 2507-2518).

[16] Singh, U., Goyal, V. and Lehal, G.S., 2012. Named Entity Recognition System for Urdu. In *COLING* (pp. 2507-2518)..

[17] Adeeba, F. and Hussain, S., 2011. Experiences in building the Urdu WordNet. *Asian Language Resources collocated with IJCNLP 2011*, p.31.

[18] Hussain, S., 2004, August. to-sound conversion for Urdu text-to-speech system. In *Proceedings of the workshop on computational approaches to Arabic script-based languages*(pp. 74-79). Association for Computational Linguistics.

[19] Ahmed, T. and Hautli, A., 2010. Developing a basic lexical resource for Urdu using Hindi WordNet. *Proceedings of CLT10, Islamabad, Pakistan*.

[20] Hussain, S. and Afzal, M., 2001. Urdu computing standards: Urdu zabta takhti (uzt) 1.01. In *Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International* (pp. 223-228). IEEE."

[21] RIAZ, M.K., RAFIQUE, M.M. and SHAHID, S.R., VOWEL INSERTION GRAMMAR.

[22] Khanam, M.H., Madhumurthy, K.V., Khudhus, M.A. and JE, B., Part-Of-Speech Tagging for Urdu in Scarce Resource: Mix Maximum Entropy Modelling System.

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

[24] Ali, S.A., Noor, M.D., Javed, M.A., Aslam, M.M. and Khan, O.A., 2016. Salience Analysis of NEWS Corpus using Heuristic Approach in Urdu Language. *International Journal of Computer Science and Network Security (IJCSNS)*, *16*(4), p.28.

[25] KHANAM, M.H., MADHUMURTHY, K. and KHUDHUS, M., Comparison of TnT, Max. Ent, CRF Taggers for Urdu Language.

[26] Mukund, S., Srihari, R. and Peterson, E., 2010. An Information-Extraction System for Urdu---A Resource-Poor Language. *ACM Transactions on Asian Language Information Processing (TALIP)*, *9*(4), p.15.

[27] Munir, S., Abbas, Q. and Jamil, B., 2017. Dependency Parsing using the URDU. KON-TB Treebank. *International Journal of Computer Applications*, *167*(12).

[28] Siddiq, S., Hussain, S., Ali, A., Malik, K. and Ali, W., 2010, December. Urdu Noun Phrase Chunking-Hybrid Approach. In *Asian Language Processing (IALP), 2010 International Conference on* (pp. 69-72). IEEE.

[29] Ali, W., Malik, M.K., Hussain, S., Siddiq, S. and Ali, A., 2010, September. Urdu noun phrase chunking: HMM based approach. In *Educational and Information Technology (ICEIT), 2010 International Conference on* (Vol. 2, pp. V2-494). IEEE.