Tagging Urdu Sentences from English POS Taggers

Adnan Naseem
Department of Computer Science
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Salman Ahmed
Department of Computer Science
International Islamic University,
Islamabad, Pakistan

Faizan Rasul Hashmi
Department of Computer Science
University of Lahore,
Lahore, Pakistan

Muazzama Anwar
Department of Computer Science
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Qadeem Akhtar Satti
Department of Computer Science
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Tahira Malik
Department of Computer Science
University of Lahore,
Lahore, Pakistan

***Abstract*—Being a global language, English has attracted a majority of researchers and academia to work on several Natural Language Processing (NLP) applications. The rest of the languages are not focused as much as English. Part-of-speech (POS) Tagging is a necessary component for several NLP applications. An accurate POS Tagger for a particular language is not easy to construct due to the diversity of that language. The global language English, POS Taggers are more focused and widely used by the researchers and academia for NLP processing. In this paper, an idea of reusing English POS Taggers for tagging non-English sentences is proposed. On exemplary basis, Urdu sentences are processed to tagged from 11 famous English POS Taggers. State-of-the-art English POS Taggers were explored from the literature, however, 11 famous POS Taggers were being input to Urdu sentences for tagging. A famous Google translator is used to translate the sentences across the languages. Data from twitter.com is extracted for evaluation perspective. Confusion matrix with kappa statistic is used to measure the accuracy of actual Vs predicted tagging. The two best English POS Taggers which tagged Urdu sentences were Stanford POS Tagger and MBSP POS Tagger with an accuracy of 96.4% and 95.7%, respectively. The system can be generalized for multi-lingual sentence tagging.**

Keywords*—Standford part-of-speech (POS) tagger; Google translator; Urdu POS tagging; kappa statistic

I. Introduction

One of the most fundamental parts of the linguistic pipeline is part-of-speech (POS) tagging. POS tagging is the process of assigning grammatical tags (nouns, verbs, adjectives, adverbs) to each word in a text. This is a basic form of syntactic analysis of the language which has many applications in NLP. Most POS taggers are trained from treebanks in the Newswire domain, such as the Wall Street Journal corpus of the Penn Treebank. However, Stanford POS Tagger is widely used by the researchers due to its multi-lingual (computer language) support packages. Such as, Docker, F#/C#/.NET, GATE, Go, Javascript (node.js), PHP, Python, Ruby, XML-RPC and Matlab. Therefore, Stanford POS Tagger is considered as an example in this paper. Output from the rest of the POS Taggers is not discussed due to the page limitations. Challenges encountered due to the termination of tagging out of domain data, and nature of Twitter text conversations, lack of traditional orthography, and 140-character length limit for each message ("Tweet").

Since, the Internet has become a major medium of social interaction and communication. Whereas, the medium of communication is English, therefore, a rich source of information pool is growing with a very fast pace comprising some useful information. However, it is a tight and hard practice to filter out the useful information from such a massive stuff. Majority of contribution regarding to developing tools took place regarding to the English based communication. In case of POS tagging a rich literature is available regarding to English POS Taggers as compared to other languages. Each POS Tagger is working decently inside its domain and within its limitations. A lot of researchers natively other than English, are also contributing in English literature. However, the valuable information other than in English language is also as important as others. Apart to bring a decent amount of researchers to take part in non-English text, an idea of reusing English tools, techniques, methodology is proposed. More specifically, English POS taggers are to be reused for tagging non English language text.

In this research, after an extensive literature review of English POS Taggers, the Stanford POS Tagger, written specifically for English sentences is reused to tag Urdu sentences as an example. Twitter API is used to extract the Urdu sentences (tweets) on a specific topic from the Twitter. After the refinement process, sample of Urdu sentences is randomly selected for further processing. Google Translator is used to translate the sampled Urdu sentences into English, for tagging from Stanford POS Taggers. The state-of-the-art English POS Taggers were extracted and included in this exercise. However, their detailed result will be included in the extended version of this study. Such English sentences were injected into the Stanford POS Tagger to yield tagged-English sentences. These tagged-English sentences are translated back to their original language with the help of Google translator. Two human annotators tagged the original sample of Urdu sentences as benchmark tagged sentences. *Kappa statistic*

along with confusion matrix is applied to measure the accuracy of each tagger for Urdu tagging.

The rest of the paper is structured as follows: Section II comprises extensive background knowledge. Section III discusses the methodology of the research. Results and Future Implications are discussed in Section IV. Conclusion, limitations and future work are placed as final sections.

II. Background Knowledge

In this section, an extensive background knowledge is presented as shown in Tables 1(a) and (b). A decent amount of literature has been carried out till date, however, current research is different in case of re-usability of benchmark POS Taggers, and generalizability of the idea. Additionally, State-of-the-Art English POS Taggers are also the part of this section.

TABLE I. (a). Background Knowledge

Sr. No	POS Tagger Name	Technique	Result	References
1	CLE Urdu Parts of Speech	CLE Urdu Digest Tagged Corpus	96.8	[1]
2	N-gram based part of speech tagger for the Urdu language	N-gram Markov Model	95.0	[2]
3	Improving part-of-speech (POS) tagging for Urdu	Humayoun's morphological analyzer, SVM Tool tagger trained	87.98	[3]
4	Solve the parts of speech tagging problem of urdu language	Hidden Markov Model		[4]
5	Four state-of-art probabilistic taggers	Tnt tagger, treetagger, RF tagger and SVM tool	95.66% by SVM tool	[5]
6	First computational part of speech tagset for Urdu	Creating one of the necessary resources for the development of a POS tagging system for Urdu		[6]
7	A rule-based methodology is used here to perform tagging in Urdu	Unitag architecture		[7]
8	NER systems for the Urdu, Hindi, Bengali, Telugu, and Oriya languages	Language specific rules and Maximum Entropy (ME)	Hindi, Bengali, Oriya, Telugu, and Urdu NER systems in terms of fmeasure were 65.13%, 65.96%, 44.65%, 18.74%, and 35.47% respectively	[8]
9	A design schema and details of a new Urdu POS tagset	The Penn Treebank	Accuracy of 96.8%.	[9]
10	Named Entity Recognition (NER) system for Urdu language	Urdu NER system		[10]
11	Named Entity Recognition	Rule-based Urdu NER algorithm		[11]
12	Problems of NER in the context of Urdu Language	IJCNLP-08 and Izaafats	Twelve NE proposed	[12]
13	NER on Conditional Random Field (CRF)	Precision, recall, and f-measure	63.72%, 62.30%, and 63.00% as values for precision, recall, and fmeasure	[13]
14	Developing a wordnet for Urdu on the basis of Hindi wordnet.	Wordnet		[14]
15	To develop models which map textual input onto phonetic content	Thus Urdu pronunciation may be modelled from Urdu text by defining fairly regular rules	Takes textual input and converts it into an annotated phonetic string.	[15]
16	With developing a lexical knowledge resource for Urdu on the basis of Hindi wordnet	Transliterators	Computational semantics based on the Urdu pargram grammar	[16]
17	UZT 1.01 standard	Unicode		[17]
18	Vowel insertion grammar for Urdu language	Building speech synthesis for Urdu language		[18]
19	Of automated Part-of-speech tagging	Maximum Entropy (ME) modelling system , Morphological analyser(MA) and stemmer	Proposed different models ME, ME+Suf, ME+MA, ME+Suf+MA	[19]
20	Release of a sizeable monolingual Urdu corpus automatically tagged with part-of-speech tags	Monolingual corpus and release the tagged corpus	88.74%	[20]
21	Analyzing the political News Corpus for finding Important Entities,	Heuristic based Salience Analysis of Urdu News Corpus	85.5	[21]

	Saliences in the Urdu language			
22	Efficient methods of computational linguistics.	Tnt tagger, Maximum Entropy tagger and CRF (Conditional Random Field)	.tnt tagger manages to obtain 93.56 for Urdu	[22]
23	Urdu-to-English transliteration	Bootstrap	84.1%	[23]
24	Evaluation of URDU.KON-TB in the dependency parsing domain.	Maltparser, The algorithm used to train and test data is Nivre arc-agear algorithm.	The experiments results show URDU.KON-TB treebank is not suitable for the dependency parsing as dependency relation because Head information was missing in the treebank.	[24]
25	Statistical model used in this work is HMM along with IOB chunk annotation	Tnt Tagger	97.52%	[25]
26	Noun phrase chunker for Urdu which is based on a statistical approach	HMM based approach	97.61	[26]

TABLE I. (b). STATE OF THE ART ENGLISH POS TAGGERS

Sr. No	**Name of POS Tagger**	**Available online?**	**Supported Programming Languages**	**Results**
1	CRF tagger	No	Java	97.00%
2	Citar - Trigram HMM part-of-speech tagger	No	C++ version available	
3	JsPOS	No	Javascript	
4	Term Extractor	No	Python package	
5	Stanford Log-linear Part-Of-Speech Tagger	No	Multiple language bindings	
6	MorphAdorner	Yes	Generic	96-97%
7	spaCy	Yes	Python/Cython	
8	SMILE Text analyzer	Yes	Java API	
9	LingPipe	No	multiple	
10	Apache OpenNLP	No	Java	
11	RDRPOSTagger	No	Python	
12	Brill's Tagger	Yes		95-97%
13	TnT	No	Multiple	95.99%
14	HunPOS	No	Multiple	95.97%
15	dTagger	No		95.1%
16	MaxEnt	No	Python, java	97.23%
17	Curran & Clark	No		97%
18	Tree Tagger	Yes	multiple	
19	Rosette based linguistic	No	Commercial Product	
20	Memory based tagger	Yes	TiMBL, C++	
21	SVM Tool	Yes but not working	SVM based	97.2%
22	ACOPOS tagger	No	C	
23	MXPOS tagger	No	Java	
24	fnTBL	No	C++ transformation based	
25	GPOSTTL	No	PHP+mysql enhanced version of brill's tagger	
26	muTBL	No	Transformation based learner	
27	YamCha	No	SVM based C/C++ open source	
28	QTag	No	HMM Java based	
29	Lingua-EN-Tagger	No	Perl	
30	CLAWS	Yes		96-97%
31	Infogistics	Yes		96-98% for known words and 88-92% for unknown words
32	AMALGAM tagger	No		
33	TATOO	No	Perl	

III. Research Methodology

This section comprises the methodology of the current research. Twitter APIs are used to extract the data on a specific topic. Data from Twitter for a novice topic PANAMA CASE is extracted with the help of Twitter API. Raw data are refined and ten sample sentences are randomly picked for further processing. Google Translator was used to translate the sampled Urdu sentences into English, for tagging from famous English POS Taggers, which were extensively explored from the literature. Such English sentences were injected into each tagger to yield tagged-English sentences. These tagged-English sentences were translated back to their original language with the help of Google translator. Two human annotators tagged the original sample of Urdu sentences as benchmark tagged sentences. *Kappa statistic* along with confusion matrix was applied to measure the accuracy of each tagger for Urdu tagging. Best two POS Tagger for Urdu sentences is hence prioritized. The whole process from step, selecting sample to find the accuracy was repeated three times to get the best results. On exemplary basis only Stanford POS Tagger is considered at this stage. The reason behind the consideration of Stanford POS Tagger here is, it outperformed the rest of the POS Taggers with 96.4% kappa statistics. The detailed results of the rest of the POS Taggers can be provided on demand. Below is the research methodology of current study in Fig. 1.

Twitter[1] is a social networking platform where millions of users communicate each day, billions of short text messages (up to 140 characters) tweets. Tweets on specific political issues were used to get tweets related to the keyword (Panama, PMLN and TTP). However, we make sure filter the unique tweets written in Urdu while we review the mesh by Twitter API[2]. To avoid re-tweets, the same check in the API is placed. The Hash functions were used to eliminate duplicate tweets. All non-Urdu characters were filtered out at the very first stage of the refinement, i.e. URLs, twitter connector (@username) and hashtags (#PTI, #PMLN) from tweets and then put them as a key in HashMap. Original tweets were used as the value of these keys. After running this procedure on all tweets, the number of tweets was reduced by approximately 40%. This remaining tweets can be safely said as unique tweets. Every Tweet was treated as a new sentence.

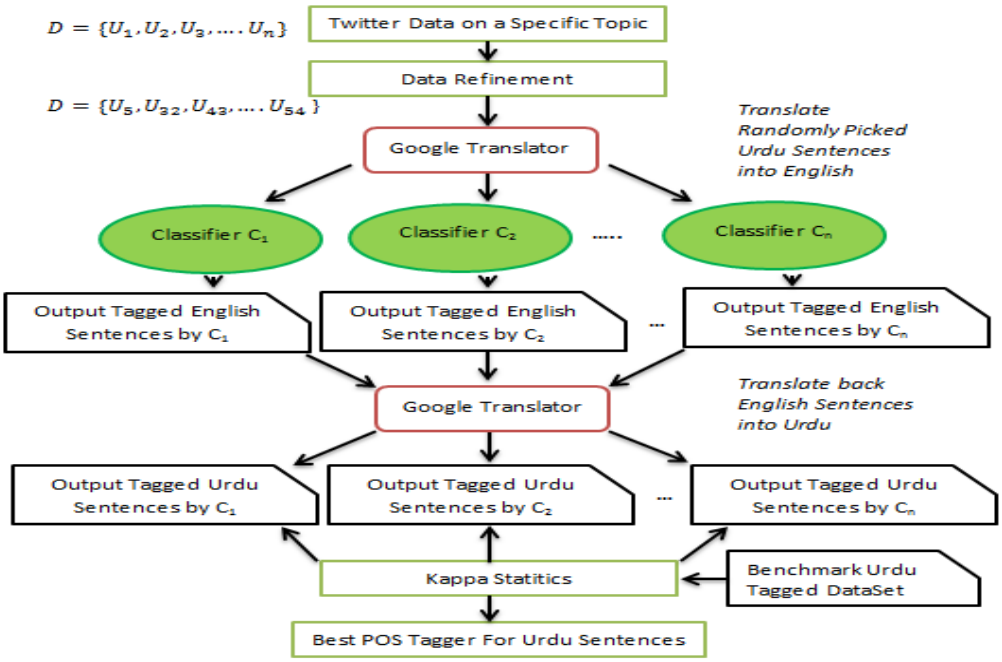

Fig. 1. Research methodology.

A random sample of 10 sentences/tweets was considered for further processing as shown in Table 2. A decent amount of literature claims different types of English POS Taggers. However, Stanford POS Tagger was used at this stage for further processing. Yet, all other state-of-the-art famous POS Taggers will be discussed extended version of current study. Moreover, these taggers can be re-useable to tag multi-lingual sentences. Additionally, the overall result of all POS Taggers is provided in Fig. 2. In order to translate sampled Urdu sentences into English sentences, an Urdu-to-English translator namely, Google Translator[3] was used.

These translated English sentences were injected into a Stanford POS tagger. The output of this step was tagged translated English sentences as resulted in Table 3.

Google translator was used again to translate back the Tagged translated English sentences into the original form, i.e. Urdu as shown in Table 4.

[1]http://twitter.com/
[2]http://twitter4j.org/en/index.html
[3]https://translate.google.com/

TABLE II. SAMPLE TWITTER SENTENCES

Sampled Urdu sentences From Twitter	S. No.
عائشہ گلالئی نے قیادت پر الزام لگایا، ضرور کچھ ہوا ہوگا۔	.(1
عوام نے پانام کیس کا فیصلہ تسلیم نہیں کیا۔	.(2
الحمد للہ آج ریلی میں 1 کروڑ لوگ شریک ہوئے۔ دیکھ سکتے ہو تو دیکھ لو۔	.(3
نوازشریف پانامہ کیس فیصلے کے بعد عوام کو گمراہ کرنے کی کوشش کررہے ہیں۔	.(4
پاناما کیس میں نااہلی کے بعد نوازشریف کا لاہور کا پہلا سفر۔	.(5
بچے کی ہلاکت پروالدین بے ہوش ہوگئے۔	.(6
نواز شریف کے قافلے میں بچہ جاں بحق۔	.(7
سابق وزیراعظم نواز شریف کا قافلہ گجرات شہرمیں داخل۔	.(8
کیپٹن ریٹائرڈ صفدر اور آصف کرمانی نے کلثوم نواز کے کاغذات نامزدگی جمع کروائے۔	.(9
آمریت کا دور اچھا ہوتا تھا سویلینز نے ملک تباہ کر دیا ہے،	.(10

TABLE III. SAMPLE TWITTER SENTENCES

Tagged English Sentences by Stanford POS Tagger	S.No
Aisha|NNP Gulalai|NNP blamed|VBD the|DT leadership|NN ,|, something|NN must|MD have|VB happened|VBN .|.	.(1
People|NNS has|VBZ not|RB recongnized|VBN panama|NN case|NN 's|POS decision|NN .|.	.(2
Today|NN ,|, there|EX are|VBP 1|CD million|CD people|NNS participating|VBG in|IN the|DT rally|NN .|. See|VB if|IN you|PRP can|MD see|VB .|.	.(3
Nawaz|NNP sharif|NN after|IN verdict|NN of|IN panama|NN case|NN is|VBZ trying|VBG to|TO mislead|VB people|NNS .|.	.(4
Nawaz|NNP Sharif|NNP 's|POS first|JJ visit|NN to|TO Lahore|NNP after|IN disqualification|NN in|IN the|DT Panama|NNP case|NN .|.	.(5
Parents|NNS became|VBD unconscious|JJ at|IN death|NN of|IN baby|NN .|.	.(6
Child|NN dies|VBZ in|IN carvan|NN of|IN nawaz|NN sharif|NN .|.	.(7
Former|JJ PM|NNP nawaz|NN sharif|NN 's|POS carvan|NN entered|VBD gujrat|JJ city|NN .|.	.(8
Captain|NN retired|VBD safdar|NN and|CC asif|NN kirmani|NNS submit|VBP nomination|NN papers|NNS of|IN kulsoom|NN nawaz|NN .|.	.(9
Dictatorship|NN was|VB good|JJ soviets|NN destroyed|VB country|NN .|PUNCT |X	.(10

TABLE IV. TAGGED URDU SENTENCES BY STANFORD POS TAGGER

Tagged Urdu Sentences by Stanford POS Tagger	S.No
عائشہ|NNP گلالئی نے|NNP قیادت|NN پر الزام لگایا|VBD، ضرور|MD کچھ|NN ہوا|VBN ہوگا|VB	.(1
عوام|NNS نے|IN پاناما|NN کیس|NN کا|POS فیصلہ|NN تسلیم|VBN نہیں|RB کیا|VBZ	.(2
الحمد للہ|RB آج|NN 1 ریلی|CD میں|CD کروڑ|CD لوگ|NNS شریک ہوئے|VBG دیکھ|VB سکتے|MD ہو تو|IN دیکھ لو|VB	.(3
نواز|JJ شریف|NN پانامہ|NN کیس|NN کی|VBG کوشش کررہے|VB کو گمراہ کرنے|NNS عوام|IN کے بعد|NN فیصلے|NN کے|IN پہلے|VBG ہیں|VBZ	.(4
سفر|NN کا|NN پہلا|NNP لاہور|POS کا|NNP شریف|NNP نواز|IN کے بعد|NN نااہلی|IN میں|NN کیس|NNP پاناما|	.(5
بچے|NN کی|IN ہلاکت|NN پر|IN والدین|NNS بے ہوش|JJ ہوگئے|VBD	.(6
نواز|NN شریف|NN کے|IN قافلے|IN میں|NN بچہ|NNP جاں بحق|VBZ	.(7
سابق|NNP وزیراعظم|NNP نواز|NN شریف|NN کا|POS قافلہ|NN گجرات|VBG شہر|NN میں داخل|VBD	.(8
کیپٹن|NNP ریٹائرڈ|VBD صفدر|VBG اور|CC آصف|NN کرمانی|NNS نے|NN نواز|IN کے|IN کاغذات|NNS نامزدگی|NN جمع کروائے|VB	.(9
ڈیکٹیٹر شپ | اچھا| NN | سوویٹس| JJ | NN | کو تباہ کر دیا| VB | ملک| NN | | PUNCT | X |	.(10

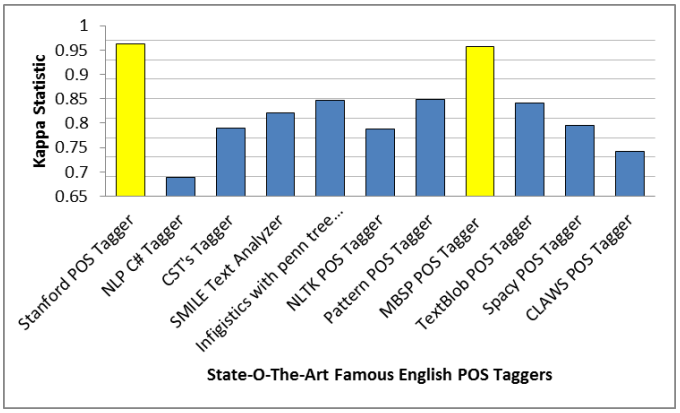

Fig. 2. Confusion matrix.

IV. Results and Future Implications

In order to check the accuracy of the subjected POS tagger with respect to Urdu language, Kappa Statistic with confusion matrix was considered. Manually annotations were applied with the help of two annotators to consider the best possible tags for original sampled Urdu data. Furthermore, Kappa Statistic with confusion matrix was applied to each tag used in Stanford POS Tagger for Urdu perspective as shown in Table 5. There were total 15 unique tags. The confusion matrix for actual tag (best possible) vs. predicted tag (tag assigned by Stanford POS Tagger) was synthesized for each of the following fifteen tags. Moreover, total accuracy and random accuracy were also calculated with the help of the following formula. Additionally, Kappa statistic was computed with the help of extracted values. The average value extracted by adding the individual kappa values of all the computed tags to the number of all tags. Accuracy of Urdu tagged sentences with the reuse of Stanford English POS Tagger was 96.4 on average, which is more than any of the existing Urdu POS Tagger. The process of randomly taking sample sentences was performed three times to remove the ambiguity of bias ness of sample selection.

Kappa Statistic

kappa= (Total accuracy - random accuracy)/ (1-random accuracy)

Total accuracy= (TP + TN)/ (TP + TN+ FP+FN)

Random Accuracy= (TN + FP)* (TN + FN) + (TP + FN)* (TP + FP)/ Total*Total

		Predicted Class	
		Not-NN	**NN**
Actual Class	**Not-NN**	TN	FN
	NN	FP	TP

Fig. 3. Confusion matrix.

In Fig. 3, TN is True Negative, FN is False Negative, FP is False Positive and TP is True Positive.

TABLE V. Kappa Statistic

			Predicted		**Total**	**Total accuracy**	**Random Accuracy**	**Kappa**	**Average Accuracy**
Tags			Not NN	NN					
		Not NN	52	0	83	0.975904	0.538104	0.947832	0.963088018
NN	**Actual**	NN	2	29					
NNP		Not NNP	74	0	83	1	0.806648	1	
	Actual	NNP	0	9					
VB		Not VB	79	0	83	1	0.90826	1	
	Actual	VB	0	4					
VBN		Not VBN	81	0	83	1	0.952969	1	
	Actual	VBN	0	2					
VBD		Not VBD	79	1	83	0.987952	0.919146	0.850987	
	Actual	VBD	0	3					
MD		Not MD	81	0	83	1	0.952969	1	
	Actual	MD	0	2					
VBG		Not VBG	81	0	83	1	0.952969	1	
	Actual	VBG	0	2					
CD		Not CD	81	0	83	1	0.952969	1	
	Actual	CD	0	2					
POS		Not POS	80	0	83	1	0.930324	1	
	Actual	POS	0	3					
NNS		Not NNS	77	0	83	1	0.865873	1	
	Actual	NNS	0	6					
RB		Not RB	82	0	83	1	0.976194	1	
	Actual	RB	0	1					
IN		Not IN	73	0	83	1	0.788068	1	
	Actual	IN	0	10					

VBZ		Not VBZ	80	0	83	1	0.930324	1	
	Actual	VBZ	0	3					
VBP		Not VBP	82	0	83	1	0.976194	1	
	Actual	VBP	0	1					
JJ		Not JJ	77	2	83	0.963415	0.896211	0.647501	
	Actual	JJ	1	2					

V. Conclusion, Limitations and Future Work

POS Tagging is considered to be an essential component of several NLP applications. The new POS Tagger is not easy to develop for unstructured data.Therefore, it affects the accuracy of tagging due to the diversity of the language. In this study, the idea of reusability of famous English POS taggers is used for tagging non-Engish sentences. A famous Google translator is used to translate the sentences across the languages. Data from twitter.com is extracted for evaluation perspective. Confusion matrix with kappa statistic is used to measure the accuracy of actual Vs predicted tagging. The result shows the accuracy of 96.4% for Stanford POS Tagger which is the best among 11 famous English POS Taggers. The system can be generalized for multi-lingual sentence tagging.

Alike other studies, current studies have also some limitations. Several translators have different translations of same sentence when translating the source language to target language. Additionally, even same translator translates a source language into targeted language, when re-translating the same text, produces different results. In this study, re-translation was carried out with the help of mapping the words. E.g. He is a boy. Wo aik larka ha. (he, wo), (aik, is), (larka, boy) and (ha, is). A customized Translator for specific language could ease the whole process. Another limitation of this study was the random selection of sentences. It was neutralized by taking the sample sentences thrice, however, the results were approximately same.

Short texts were used in this study; however, text other than from twitter will be used in an upcoming paper. Apart from the overall results, a detailed comparison of state-of-the-art English POS Taggers will be considered to rank the best POS Tagger for Urdu sentence tagging in the near future. Furthermore, sample data other than twitter will be considered for validation purposes. The current methodology could be used to tag multi-lingual tagging for the extraction of useful information. Therefore, a generic methodology for several different languages will be considered in future. Additionally, each language has different level of diversity; therefore, same methodology could be applied to several languages to avoid the development of novice complex taggers.

References

[1] Adeeba, F., Akram, Q., Khalid, H. and Hussain, S. “CLE Urdu Books N-grams”, poster presentation in Conference on Language and Technology ,(CLT 14), Karachi, Pakistan, 2014.

[2] W. Anwar, X. Wang, L. Li, and X. L. Wang, “A statistical based part of speech tagger for urdu language,” Proc. Sixth Int. Conf. Mach. Learn. Cybern. ICMLC 2007, vol. 6, no. August, pp. 3418–3424, 2007.

[3] B. Jawaid and O. Bojar, “Tagger Voting for Urdu,” Proc. Work. South Southeast Asian Nat. Lang. Process. Coling 2012, no. December 2012, pp. 135–144, 2012.

[4] Anwar W. Anwar, W., Wang, X., Lu-Li, “Hidden markov model based part of speech tagger for urdu.,” Information Technology Journal, vol.6, no.8, pp.1190-1198, 2015.

[5] H. Sajjad and H. Schmid, “Tagging Urdu Text with Parts of Speech : A Tagger Comparison,” Proc. 12th Conf. Eur. Chapter ACL, EACL’09, no. April, pp. 692–700, 2009.

[6] A. Hardie, “Developing a tagset for automated part-of-speech tagging in Urdu,” Corpus Linguist., pp. 1–11, 2003.

[7] A. Hardie, “The computational analysis of morphosyntactic categories in Urdu,” PhD diss., Lancaster University, 2004 .

[8] S. Chatterji, “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. 2008.

[9] T. Ahmed et al., “The CLE Urdu POS Tagset.” In LREC 2014, Ninth International Conference on Language Resources and Evaluation, pp. 2920-2925. 2015.

[10] S. Naz, A. Iqbal Umar, S. Hamad Shirazi, S. Ahmad Khan, I. Ahmed, and A. Ali Khan, “Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language,” Res. J. Appl. Sci. Eng. Technology., vol. 8, no. 10, pp. 1272–1278, 2014.

[11] K. Riaz, “Rule-based Named Entity Recognition in Urdu,” In Proceedings of the 2010 named entities workshop, Association for Computational Linguistics, pp. 126-135, 2010.

[12] U. Singh, V. Goyal, and G. Singh Lehal, “Named Entity Recognition System for Urdu,” In COLING, pp. 2507–2518, 2012.

[13] M. K. Malik and S. M. Sarwar, “Urdu Named Entity Recognition And Classification System Using Conditional Random Field,” Sci.Int.(Lahore), vol. 27, no. 5, pp. 4473–4477, 2015.

[14] F. Adeeba and S. Hussain, “Experiences in building the Urdu WordNet,” Asian Language Resources collocated with IJCNLP 2011, vol. 13, pp. 31–35, 2011.

[15] S. Hussain, “Letter-to-Sound Conversion for Urdu Text-to-Speech System.” In Proceedings of the workshop on computational approaches to Arabic script-based languages, Association for Computational Linguistics, pp. 74-79. 2004.

[16] T. Ahmed and A. Hautli, “Developing a Basic Lexical Resource for Urdu Using Hindi WordNet.” Proceedings of CLT10, Islamabad, Pakistan, 2010.

[17] S. Hussain and M. Afzal, “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, 2001.

[18] M. Khurram Riaz, M. Mustafa Rafique, and S. Raza Shahid, “Vowel Insertion Grammar.”

[19] M. Humera Khanam, K. V Madhumurthy, A. Khudhus, and A. Professor, “Part-Of-Speech Tagging for Urdu in Scarce Resource: Mix Maximum Entropy Modelling System,” Int. J. Adv. Res. Comput. Commun. Eng., vol. 2, no. 9, 2013.

[20] B. Jawaid, A. Kamran, and O. Bojar, “A Tagged Corpus and a Tagger for Urdu.” In LREC, pp. 2938-2943. 2014.

[21] S. A. Ali et al., “Salience Analysis of NEWS Corpus using Heuristic Approach in Urdu Language,” IJCSNS Int. J. Comput. Sci. Netw. Secur., vol. 16, no. 4, 2016.

[22] M. Humera Khanam, K. V Madhumurthy, and A. Khudhus, "Comparison of TnT, Max.Ent, CRF Taggers for Urdu Language," Int. J. Eng. Sci. Res., vol. 4, no. 1, 2013.

[23] S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu—A Resource-Poor Language," ACM Trans. Asian Lang. Inf. Process. ACM Ref. Format ACM Trans. Asian Lang. Inform. Process, vol. 9, no. 4, pp. 15–43, 2010.

[24] S. Munir, Q. Abbas, and B. Jamil, "Dependency Parsing using the URDU.KON-TB Treebank," Int. J. Comput. Appl., vol. 167, no. 12, pp. 975–8887, 2017.

[25] S. Siddiq, S. Hussain, A. Ali, K. Malik, and W. Ali, "Urdu Noun Phrase Chunking - Hybrid Approach," in 2010 International Conference on Asian Language Processing, pp. 69–72, 2010.

[26] W. Ali, M. Kamran Malik, S. Hussain, S. Siddiq, and A. Ali, "Urdu noun phrase chunking: HMM based approach," in 2010 International Conference on Educational and Information Technology, 2010.