EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Supervised Urdu Word Segmentation Model Based on POS Information

Sadiq Nawaz Khan[1,*], Khairullah Khan[1] and Wahab Khan[2]

[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[1]Department of Computer Science, University of Science & Technology Bannu, Pakistan
[2]Department of Computer Science & Software Engineering, IIU, Islamabad 44000, Pakistan

Abstract

Urdu is the national language of Pakistan, also the most widely spoken and understandable language of the globe. In order to accomplish successful Urdu NLP a robust and high-performance NLP tools and resources are utmost necessary. Word segmentation takes on an authoritative role for morphologically rich languages such as Urdu for diverse NLP domains such as named entity recognition, sentiment analysis, part of speech tagging, information retrieval etc. The morphological richness property of Urdu adds to the challenges of the word segmentation task, because a single word can be composed of null or a few prefixes, a stem and null or a few suffixes. In this paper we present supervised Urdu word segmentation scheme based on part of speech (POS) information of the corresponding words. For experiments conditional random fields (CRF) with contextual feature is used. The performance of the proposed system is evaluated on 300K words, results shows evidential improvements on baseline approach.

Keywords: Urdu, Word segmentation, supervised learning, conditional random fields

Received on 10 May 2018, accepted on 04 September 2018, published on 10 September 2018

doi: 10.4108/eai.19-6-2018.155444

*Corresponding author. Sadiqnawaz97@gmail.com

1. Introduction

Nowadays Natural Language Processing plays a vital role in every field of computer science. Human beings are trying to simulate human knowledge by computer system. For this purpose, NLP researchers struggle by introducing knowledge through which computers understand and use natural language. To achieve desired tasks different types of advanced tools and procedures are applied to make computer systems more cognizable. Various disciplines lie in NLP fundamentals such as electronic and electrical engineering, linguistics, information and computer sciences, mathematics, psychology and artificial intelligence (AI) etc [1]. Natural Language Processing applications are widely used which mainly consist of different fields of studies, like word segmentation, speech recognition, text processing and summarization, CLIR (cross language information retrieval), user interfaces, voice recognition and artificial intelligence etc. Information retrieval (IR) recognizes desired valuable information from a huge collection of data while information extraction (IE) is used to process document(s) for identification of such entities or events that are pre-specified or a technique that processes a document(s), to identify pre-specified entities or events. Artificial intelligence is a sub-field of computer science in which we study the development of hardware and software that simulates human intelligence.

For every NLP application Word Segmentation has vital role. Word segmentation is capable of separation written or oral text into meaningful word tokens. It identified words boundaries in a spoken language. In

recent age, Hindi like languages attracted researcher's attention more than other regional languages. For example regional language Urdu is becoming popular on web day by day[2]. Data Mining (DM) and Informational retrieval (IR) demands for exploration of natural language processing with responsibilities of the topic categorization, relationship exploration, sentiment analysis and event extraction etc. NLP significant tasks i.e NER (named entity recognition), POS (part-of-speech) tagging, stop words removal, morphological analysis, parsing have major role in all NLP systems [3].

In case of Urdu language (one of the most of important language of the Asian countries) segmentation is much difficult as some of the other Asian languages, in which space is used for word boundary. Urdu word segmentation issues i.e space insertion and omission are caused by use of space in Urdu text [4] and [5]. The Space omission problem in Urdu word segmentation e.g. the Urdu word آپکا(Aapka, Your's) which is composed of two words آپ(Aap, You) and کا(Kaa, of) but for the system it will be treated as a single word. Segmentation of such like Urdu words have been addressed by [6] using Urdu-Devnagri transliteration system. The second problem i.e Space Insertion problem consider the Urdu word ضرورت مند (Zaroratmand, Needful) which is a single word but the system will consider it as two words while segmentation i.e ضرورت (Zarorat, need) and مند (mand), such types of Urdu words have been handled by [7] using two-stage system. [8] and [9] have briefly discussed Hindi-Urdu transliteration issues while doing segmentation. Segmentation for Sindi language using simple, compound and complex words are discussed in three layers [10]. Urdu-Arabic Word Segmentation techniques and also their challenges has been discussed by [11].

As for our knowledge up to now, there has no research work conducted from the ULP research community in which the researchers have examined the effect of supervised machine learning model especially CRFs and POS information for Urdu word segmentation task.

Therefore, this gap and advantages of supervised learning schemes in diverse areas, motivated us to undertake supervised learning model along with POS tag information for Urdu word segmentation problem.

The main contributions of this study are below:

(1) It is the first to employ CRF along with POS information as feature for the subject ;

(2) It is a systematic evaluation of CRF models on data of three different genre such as International, sport and science news

2. Literature review

Now a day's different languages use different techniques for word segmentation problem so far. These techniques are used by NLP researchers and have deduced better results from each one. The existing techniques for word segmentation in NLP are Dictionary/rule based, statistical/machine learning and hybrid approaches.

2.1. Existing techniques

The segmentation techniques used so far, are classified into below categories:

- Dictionary/Rule based techniques
- Statistical/Machine learning based techniques
- Hybrid techniques
- Feature based techniques

2.1.1. Dictionary/ Rule based techniques

To perform different Urdu Language Processing tasks, Rule based techniques are highly examined in case of Urdu word segmentation so far. In these techniques, set of rules or pattern are used to perform significant NLP tasks. But there are number of limitations of these techniques because for each process there will be a separate rule. These techniques were used for chines word segmentation [12]. Word segmentation in Japanese, Chines, Thai and Urdu etc are more difficult from western languages like English, French etc, because of non-formal use of space. The Urdu language is also a resource poor language. Thai word segmentation using rule based approach has been presented by [13]. The Urdu stemmer "Assas-Band" remove prefix first from stem, then remove postfix and finally stem is extracted [14]. Its accuracy rate is 91.2%. The handwritten information are converted into their corresponding Urdu text using Urdu online handwriting recognition system using these techniques[15]. NER for automated text processing in Urdu using rule based algorithm [16]. Rule based maximum matching approach for space insertion and omission problems in Urdu word segmentation have been addressed by [17].

2.1.2. Statistical/ Machine learning techniques

In recent era of research, statistical techniques surplus other techniques. Machine learning techniques giver better results than rule-based approaches, however not proper attention has been given to these techniques in case of Urdu word segmentation. ML algorithms are capable of defining a function that take input samples to a range of output values. For these algorithms a corpus is constructed in which word boundaries are explicitly defined. Statistical models concerned with bi and tri-gram models are more frequently employed. In natural language processing, supervised statistical learning most significant technique. Supervised statistical approach induces rules from training data automatically. ML algorithms comprise of intelligent modules and different ML models have been addressed by [18]. The space

insertion issue in Urdu word segmentation has been handled using statistical based technique [7] while space omission has been addressed by [6] using the same approach.

2.1.3. Hybrid techniques

These approaches combined the features of both rules based and statistical techniques. In Urdu language, Sentence boundary identification is initial step for ULP tasks which has been presented by [19] using hybrid approach along with unigram statistical model and rule based algorithm. The result was calculated better than both rule base and statistical approaches. Hybrid technique for segmentation presented by [20] for line segmentation of Urdu text using top down mechanism and for segmentation of line into ligatures using bottom up design. The desired results are quite reasonable with accuracy achieved 99.2%.

2.1.4. Feature based techniques

These techniques are used for identification word boundaries. Any specific information regarding undefined words can be tested using feature. For Thai word segmentation, features can be automatically extracted from training corpus using ML algorithm Winnow [21].

3. The Urdu language

Urdu is one of important language in Asian countries. It is the Pakistan's National language. The hand-held devices like mobiles phones etc are with success mistreatment all over however the code they supply for user input is generally in English and in Asian nation it's tough for a standard man to speak in English simply. In order to facilitate Urdu speakers and author and cut back the distinction between the common person and therefore the new technology Urdu information processing systems area unit needed. Our struggle is for reducing this gape using ML techniques for segmenting Urdu text.

Urdu is written in Arabic script which is much cursive. Arabic language has many writing styles, one of which is Nastaleeq (character based, right to left style) is commonly used for Urdu language. But NLP tasks for Arabic language are not applicable to Urdu such like stemming algorithm for Arabic will not work for Urdu. Urdu characters have some unique features i.e joining and non-joining. Joiner characters are joined with prefix and postfix characters and non-joiner characters can join only with prefix character not with postfix characters. Table 1 shows Urdu Numbers and characters. Despite of these, diacritics, punctuation marks, signs & symbols are also used in Urdu text shown below in Table 2. The joiners and non-joiners characters are listed in Table 3.

Table 1. Urdu digits and alphabets

Urdu Characters	Urdu characters	
	Numbers	Characters
	٤ ٥ ٦ ٧ ٨ ٩ ٠ ١ ٢ ٣	ا ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ر ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل م ن و ہ ی ے

Table 2. Diacritics, punctuations & symbols

Urdu Diacritics	Urdu diacritics		
	Characters not counted as part of alphabet/diacritics	Punctuation marks	Sign & Symbols
	ء ؤ ئ ۓ ۂ آ ں ھ [illegible]	، , ۔ ؟ ؛ ،	ع ﷺ [illegible] بِسمِ

Table 3. Joiners & non-Joiners

Joiners	Non-Joiners
ب ت ۃ ث ٹ ج چ ح خ ہ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ء ی	ا آ د ڈ ذ ر ز ڑ ژ و ے

3.1. Morphological richness of Urdu text

In Urdu, for each word there exists various variants, that's why it is morphological rich. Urdu is sort of the same as different Indo-European languages, e.g. for morphology, having concatenative inflective morphological system. However, in some cases, variations are often found just in case of causative verbs that conjointly exhibit stem-internal changes. Urdu morphology and its solution is presented by [22].

3.1.1. The Urdu verbs

Verb (**فعل**) plays a vital role in every language because it convey action. ULP researchers should learn the Urdu verbs because its structure is used in everyday conversation. Below Table 4 shows some examples of Urdu verbs.

Table 4. Urdu Verbs

Verbs used in Urdu		
Verbs (Urdu)	Roman Urdu	English verbs
لکھنا	Likhna	Write
مطالعہ کرنا	Mutala karna	Study
مسکرایا	Muskuraya	Laughed
لینا	lena	Take

3.1.2. The Urdu nouns

Noun (**اسم**) represents the name of place, person, animal or anything. Urdu nouns are classified into different categories and types [23]. The Table 5 consists of different types of Urdu nouns and their examples.

Table 5. Urdu Nouns

Urdu Nouns	Nouns used in Urdu		
	Type of Nouns (Urdu)	English	Example
a)	اسم عرفہ	a) Proper noun	پکلستان
i.	خطاب	i. Title	قطب اعظم
ii.	لقب	ii. Attributive	کلھم اللہ
iii.	تخلص	iii. Nom-de-plume	جالی
b)	اسم نکرہ	b) Common noun	لیک
i.	کیفیت	i. State	خوشی
ii.	جمع	ii. Collective	ٹیم
iii.	ظرف	iii. Locative	وقتگھر

3.2. Urdu linguistic resources

Urdu lexical resources square measure necessary a part of each informatics system for Urdu language. Those study concerning linguistics in Asian nation typically been restricted to those fields concerning connected Linguistics, significantly English and socio-linguistics. Next to no value of effort want been completed within description and theoretic linguistic etymology. In Pakistan there is a just as set limit in this range.

In recent era, two choices during creation of datasets for scripting languages (i.e Arabic etc) are widely used, which are: (a) Unicode character set (b) XML file format. These two choices are used for storing data. Urdu uses Unicode encoding scheme for storage however Urdu text can be stored in different kinds of format like '.txt', 'inp', or 'doc' but XML is more adaptable one because data in this format can be easily converted to other formats and easily readable for user and system as well [18].

Urdu language processing researchers are trying to develop advanced tools and datasets. The available datasets are not in good number to meet all criteria of ULP tasks. There are three ULP datasets are available currently. The detail of these dataset are given in the Table 6. The Backer-Riaz dataset first introduced by [24]. The EMILLE dataset for ULP was released by Lancaster University in 2003 [25] . The CLE (Center for Language Engineering) in Pakistan are trying for building ULP corpus. They have launched Urdu Digest POS Tagged datasets but are not freely available to ULP researcher community. The IJCNLP-2008 NE tagged dataset was annotated by "Center for Research in Urdu Language Processing" at National University of Computer & Engineering Sciences in Pakistan and IIIT Hyderabad (India). This corpus is freely available on URL http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5 [26]. While the "UNER Dataset" in ULP research community has been introduced by [27].

Table 6. Summary of existing Datasets for ULP

Name of Dataset	Year	Tasks	Words	Documents
Backer-Riaz Dataset	2002	Stop-words removal, IR, NER, stemming, baseline evaluation of Urdu & Hindi	20,000 – 50,000	7000
EMILLE Dataset	2003	Segmentation, NER, stop-words removal, translating, anaphoric annotation, language engineering analysis, POS tagging.	1,640,300 (written Urdu text) 512,000 (spoken text)	
IJCNLP-2008 NE tagged datasets	2008	NER	40,000	Training 5 Documents, Testing 1 Document
N-gram based Urdu NER tagged dataset	2012	NER	31,860	4 entity classes
URDU NER Dataset	2016	NER	48673	National News: 60 Sports News: 50 International News: 40

4. Conditional Random Fields

Conditional Random Fields (CRF) is a ML algorithm, used for various NLP tasks such as Name Entity Recognition, sequential labelling and word segmentation etc. These are undirected graphical models which are used to calculate the conditional probability of values on designated output nodes given values on designed input nodes. Comparing to HMM, CRF results better [28]. CRF can be defined using variable X and Y as:

Let the graph $G = (V, E)$ such that $Y = (Y_y)$vεv) so that Y is indices by the vertices of G. Then (X, Y) is conditional random field when the random variable Y_v, conditioned on X, obey the Markov property with respect to the graph: $p\,(Y_y/X, Y_{W,}\, w{\sim}v)$ means that w and v are neighbors in G. For sequence tagging tasks, the LDCRF (Latent-dynamic random fields) or DPLVM (Discriminative Probabilistic Latent Variable Models) are a type of CRFs for sequence tagging tasks. These models are known as latent variable models that are trained discriminatively. According to LDCRF let a given sequence of observations say, $X = x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n$ one of the tagging task but here the problem arises that how to assign sequence of labels and this problem should be solved by the model let $Y = y_1, y_2, y_3, \ldots\ldots\ldots\ldots y_n$, be a labels sequence. In ordinary linear-chain CRF, latent variables 'h' is inserted between x and y rather than directly modeling $P(Y/X)$. it uses chain rule probability.

$$P(Y/X) = \sum_h p(Y/h, X)P(h/X) \quad (1)$$

Suppose $x_{1:n}$ is a sequence of Urdu words in a sentence with name entities $z_{1:n}$. According to linear chain CRF the conditional probability is as:

$$P\,(z_{y:n}/x_{1:n}) = 1/Z \exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F} \lambda_i\, f_i(z_{n-1}, z_n, z_{1:n}, n)\right) \quad (2)$$

Where the normalization factor Z is calculated as under

$$Z = \sum_{z_{1:n}} \exp\left(\sum_{n=1}^{N}\sum_{i=1}^{F} \lambda_i\, f_i(z_{n-1}, z_n, z_{1:n}, n)\right) \quad (3)$$

For ULP tasks, many experiments have done for named entity recognition using minor use of CRF, but nowadays need sophisticated researcher's attention while using CRF as a module for NER in Urdu.

CRFClassifier gives a general usage of (self-assertive request) straight chain CRF arrangement models for any assignment. In case of proposed work, the NER structured as, e.g the given Urdu sentence:

والدین کا احترام ہمارا فرض ہے۔

Parent's respect is our duty.

[والدین:PER] [کا:PREP] [احترام:ADJ] [ہمارا:NOUN] [فرض:VERB] [ہے:PREP].

5. Issues in Urdu word segmentation

Compare to Western languages like English, French etc, Urdu faces a lot of segmentation issues because of no regular use of space in between the Urdu words. Urdu faces Space omission issue, space insertion issue, reduplicated words, compound words, affixations and English Abbreviations. These Urdu word segmentation issues are discussed below in detail:

5.1. Space insertion issue

This problem arises during word segmentation when a space is inserted in between two Urdu words. No space is written between two words in hand written Urdu text [8], [17] and [6]. While typing in computer system a space must be entered if the last character of the word is joiner, otherwise it will be joined to the next word and form incorrect/miss-understandable form and become difficult to recognize for system as well as for own native speaker. The Space insertion issue arises due to joiner characters of Urdu text [17]. Examples of space insertion issues are briefly discussed in Table 7.

5.2. Space omission issue

This problem arises in a case if we omit space in such a place where space is must to insert in between two Urdu words. Space omission in Urdu text is also a challenging task for word segmentation. But in case of joiner as a last character, space should be inserted after joiner otherwise it will append to next word and give visually false form considerable work for handling space omission issue has been done by [6]. This work consists of transliteration system from Urdu to Devnagri for segmentation and then from Devnagri to Urdu. Examples of space omission issue in Urdu word segmentation is discussed in detail in Table 7.

5.3. Compound words

These words are combination of two or more words. These combined lexemes form another lexeme [27]. Compound words are categorized into three different formats [8] which are listed in Table 7 with examples.

5.4. Reduplicated Urdu words

The words that occur twice one after each are known as reduplicated words. Urdu reduplicated words are briefly discussed by [8]. Such kind of words make Urdu word segmentation challengeable. Examples are discussed in Table 7.

5.5. The Urdu Affixations

By Affixation we mean that such words that contain prefixes and suffixes . Urdu text also contains affixation that make the segmentation process challengeable. These words are discussed by [3]. Prefixes and suffixes are removed in stemming. Examples of such Urdu words are shown in Table 7.

5.6. Abbreviations

In Urdu language, words from other languages i.e English, Arabic, Greek, Farsi, Latin etc. English abbreviations in Urdu writing need a space/dash character in between the words [8]. Some examples of abbreviations are discussed in Table 7.

Table 7. Summary of Urdu word segmentation issues

Urdu word segmentation challenges			
Issue	Form A	Form B	English translation
Space insertion	خبردار	خبر دار	Attention
Space Omission	معلوماتی	معلومات	Scientific Information
Compound Words	ماں باپ(Maa Baap), شان و شوکت(Shaan o Shaukat), موسم بہار(Mosam e Bahar)		Parents, Glory, Spring Season
Reduplicated	بار بار(Baar Baar)		Over and over again
Affixations	بدسلوک(Bad Salook), عزت دار(Ezzat Dar)		Bad behaviour, Honorable
Abbreviations	ایل ایل ایم، بی بی سی، پی آئی اے		L.L.M, BBC, PIA

6. Proposed supervised model

Our proposed system for Urdu word segmentation model based on Part-of-Speech tagging uses named entities with POS of each Urdu word as a feature and gives desired result.

We have legally used CLE POS tagged corpus for POS tag information and used UNER dataset [27] for named entity information. UNER dataset contains only named entity tags, the structure of the UNER dataset is shown in below:

<DESIGNATION>کپتان</DESIGNATION> سمیت ون
سری <LOCATION> ڈے ٹیم میں شامل کئی کے ڑی پلے ہی
میں موجود ہیں، جگر ٹیم خب پلیئرز </LOCATION>لنکا
<NUMBER>5</NUMBER> روزہ کیمپ میں نمازتراویح
کے بعد ٹریننگ کی کلاس گی رمضان الہارک کی وجہ سے کیمپ
کا شیڈول معمول سے ہٹ کر ہوگا

For feature learning POS tags are assigned to each word of UNER dataset and then used these features for subject task. Longest maximum matching technique is used for assigning Part-of-Speech tags to words.

When POS tags are assigned to UNER dataset then CRF model is trained on this UNER dataset which contains both POS and NE tags. Now this dataset is used for generating model file. The lexical dictionary file is used for testing data and is matched with this dataset. The output file is generated and results are calculated.

The proposed CRF based Urdu word segmentation model makes use of named entities and POS information of words as feature for the subject task.

Main feature used in train of CRF are given below:

- Current word
- Left context word of the current word
- Right context word of the current word
- Joint use of current word and its POS tag
- Current word and POS tag of N-1 word
- Current word and POS tag of N+2 word

7. Experiments and results

We have used WordSeg †libraries a C# implementation for evaluation of our proposed system performance. Our training corpus consists of about 300K words. Performance of the system is evaluated using Precision, Recall and F-score. Precision is the ratio between correct segmentation and total number of words to be segmented while Recall is the ratio of correct segmentation and correct plus false segmentations. F-score is the harmonic mean of Precision and Recall. We take Urdu text from ‡BBC site from three domains i.e international, sports and science. The text is in form of sentences and 3 sentences have been from each domain. The Precision Recall and F-score values for the tested text are shown in Table 8 while **Figure 1** depicts the same results graphically.

Table 8. Precision, Recall & F-score values of tested text

Tested Text		Precision, Recall & F-score		
Domain	Words	Precision	Recall	F-score
International	50	94%	51%	66%
Sports	49	96%	51%	66.6%
Science	53	98%	50.4%	66.5%

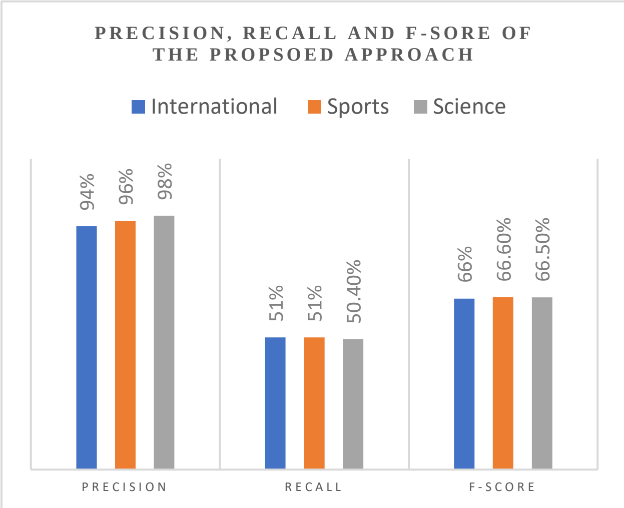

Figure 1:Graphical representation of Precision, Recall and F-Score of the proposed approach

The results show the average values of Precision, Recall and F-score for all the tested four cases are 96%, 50.8% and 66.3% respectively. By increasing the training data, the proposed system gives better results and improves efficiency. The accuracy of our system is 96%. Our system covers almost all the challenges facing to Urdu word segmentation space omission & insertion problems, compound words and reduplication foreign words.

Our proposed work has been compared with [5], in which rule based techniques using maximum matching approach is used. Before segmentation of Urdu text, this system removes diacritics from Urdu words and perform segmentation. Table 9 shows the comparison of our proposed system with baseline work:

Table 9. Comparison of proposed CRF model with baseline approach

† https://github.com/zhongkaifu/CRFSharp

‡ http://www.bbc.com/urdu/sport

Approach	Problem Addressed	Tested text	Segmented	Not segmented	Accuracy
Baseline	Space omission	2367	2209	158	93.3%
CRF	Space insertion, omission, compound, reduplicated, Abbreviations & reduplicated words	152	146	06	96%

8. Conclusions

In this paper we have presented a supervised machine learning scheme for solving Urdu word segmentation. As Urdu is much cursive language as compare to other Asian languages because of space problems in between the words. Since Urdu word segmentation is a challenging task due to its rich morphological richness property. Therefore, it requires standard tools to dealt with. In this study we tried to handle the subject problem with supervised machine learning model namely CRF along with words corresponding POS information. This is the first time that such a segmentation scheme for handling space insertion, space deletion, compound words and reduplicated words has been presented. Results shows evidential improvements of the proposed scheme over the previous approaches. In future we are planning to test deep learning approaches such deep convolutional neural networks, recurrent neural networks and LSTM networks for the subject task.

References

[1] G. G. Chowdhury, "Natural language processing," *Annual review of information science and technology*, vol. 37, pp. 51-89, 2003.

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

[3] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," *Artificial Intelligence Review*, pp. 1-33, 2016.

[4] N. Durrani, "Typology of word and automatic word Segmentation in Urdu text corpus," 2007.

[5] N. Durani and S. Hussain, "Urdu Word Segmentation, Human Language Technologies," in *The Annual Conference of the North American Chapter of the ACL, Los Angeles, California*, 2010, pp. 528-536.

[6] G. S. Lehal, "A word segmentation system for handling space omission problem in Urdu script," in *23rd International Conference on Computational Linguistics*, 2010, p. 43.

[7] G. S. Lehal, "A two stage word segmentation system for handling space insertion problem in Urdu script," *analysis*, vol. 6, p. 7, 2009.

[8] B. Jawaid and T. Ahmed, "Hindi to Urdu conversion: beyond simple transliteration," in *Conference on Language and Technology*, 2009.

[9] A. Malik, L. Besacier, C. Boitet, and P. Bhattacharyya, "A hybrid model for Urdu Hindi transliteration," in *Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration*, 2009, pp. 177-185.

[10] J. Mahar, H. Shaikh, and G. Memon, "A Model for Sindhi Text Segmentation into Word Tokens," *Sindh University Research Journal-SURJ (Science Series)*, vol. 44, 2012.

[11] A. Mahmood, "Arabic & Urdu Text Segmentation Challenges & Techniques," vol. IV, pp. 32-34, 2013.

[12] D. D. Palmer, "A trainable rule-based algorithm for word segmentation," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 1997, pp. 321-328.

[13] Y. El Hadj, I. Al-Sughayeir, and A. Al-Ansari, "Arabic part-of-speech tagging using the sentence structure," in *Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt*, 2009.

[14] Q.-u.-A. Akram, A. Naseer, and S. Hussain, "Assas-Band, an affix-exception-list based Urdu stemmer," in *Proceedings of the 7th workshop on Asian language resources*, 2009, pp. 40-46.

[15] S. Malik and S. A. Khan, "Urdu online handwriting recognition," in *Emerging Technologies, 2005. Proceedings of the IEEE Symposium on*, 2005, pp. 27-31.

[16] K. Riaz, "Rule-based named entity recognition in Urdu," in *Proceedings of the 2010 named entities workshop*, 2010, pp. 126-135.

[17] N. Durrani and S. Hussain, "Urdu word segmentation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 528-536.

[18] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," *Kuwait Journal of Science*, vol. 43, 2016.

[19] Z. Rehman and W. Anwar, "A hybrid approach for urdu sentence boundary disambiguation," *Int. Arab J. Inf. Technol.*, vol. 9, pp. 250-255, 2012.

[20] G. S. Lehal, "Ligature segmentation for Urdu OCR," in *2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 1130-1134.

[21] P. Charoenpornsawat, B. Kijsirikul, and S. Meknavin, "Feature-based thai unknown word boundary identification using winnow," in *Circuits and Systems, 1998. IEEE APCCAS 1998. The 1998 IEEE Asia-Pacific Conference on*, 1998, pp. 547-550.

[22] M. Humayoun, H. Hammarström, and A. Ranta, *Urdu morphology, orthography and lexicon extraction*: Chalmers tekniska högskola, 2006.

[23] A. Ali, S. Hussain, K. Malik, and S. Siddiq, "Study of Noun Phrase in Urdu," 2007.

[24] D. Becker and K. Riaz, "A study in urdu corpus construction," in *Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12*, 2002, pp. 1-5.

[25] P. Baker, A. Hardie, T. McEnery, and B. Jayaram, "Corpus data for South Asian language processing," in *Proceedings of the 10th Annual Workshop for South Asian Language Processing, EACL*, 2003.

[26] S. Hussain, "Resources for Urdu Language Processing," in *IJCNLP*, 2008, pp. 99-100.

[27] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "Named Entity Dataset for Urdu Named Entity Recognition Task," *Organization*, vol. 48, p. 282, 2016.

[28] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the eighteenth international conference on machine learning, ICML*, 2001, pp. 282-289.