University of Sindh Journal of Information and Communication Technology (USJICT)
Volume 4, Issue 3, October 2020
ISSN-E: 2523-1235, ISSN-P: 2521-5582
Website: http://sujo.usindh.edu.pk/index.php/USJICT/

Extracting Temporal Entity from Urdu Language Text

Daler Ali[1], Malik Muhammad Saad Missen[1], Muhammad Ali Memon[2], Muhammad Ali Nizamani[2], Asadullah Shaikh[3]

Department of CS and IT, Isalamia University, Bahwalpur
IICT, University of Sindh, Jamshoro
College of Computer Science and Information System, Najran University
idalerali@gmail.com, saad.missen@iub.edu.pk, ma.nizamani@usindh.edu.pk

Abstract: The detection of temporal entities within natural language texts is an interesting information extraction problem. Temporal entities help to estimate authorship dates, enhance information retrieval capabilities, detect and track topics in news articles, and augment electronic news reader experience. Research has been performed on the detection, normalization and annotation guidelines for English temporal entities. However, research for Urdu language lags far behind and there is a need for lot of work to be done in this regard especially when huge quantity of Urdu data is being generated on online social networks on daily basis. In this paper, we propose a rule-based approach for temporal entity extraction for Urdu language. Comparing our approach with existing Urdu temporal entity extraction approaches, our approach dominates on behalf of accuracy and on tackling with all types of Urdu temporal entity types. We use a publicly available Urdu data corpus for our experiments which consists of 206 date tags. We extend this dataset by adding 200 Urdu Fully Qualified Date (UFQD) tags. We also introduce a new date type for Urdu language called Urdu Partially Fully Qualified. Our proposed system achieved average (0.97, 0.98 and 0.98) (Precision, Recall and F1-Measure) respectively for UFQD and Urdu Partially Fully Qualified Date. Some challenges and issues of other date types in Urdu Textual Language i.e. Deictic and Anaphoric are also discussed in detail.

Keywords: Entity Extraction, Urdu Language Text, Dates

I. INTRODUCTION

With the evolutionary growth in number of Internet users, huge volumes of structured and unstructured data are being generated on daily basis [1]. One of the other reasons behind this exponential increase in user generated data is emergence of online social networks. Usage of online social media is finding new trends online. For example, according to a report 88% Indians are non-English speakers while 60% Indians in urban areas access online content in Hindi, Tamil and Marathi languages. Similar trends can be traced for almost all local languages in relevant population. Urdu is one the most spoken language in the world [2]. It is also National language of Pakistan. With 120 million mobile phone users and availability of Internet on cheap rates, Urdu is becoming the mainstream language of communication online among its speakers. Similarly, Google provides input tools for different languages. With these significant statistics, a number of Urdu writing tools have been made available online that can also be considered one of the reasons Urdu data is increasing online. Fr example, Google input tool for Urdu language provides fast and easy typing using virtual keyboard and typing in roman Urdu. Windows operating system is also facilitating users to select regional language for management of electronic devices i.e. laptop and mobile phones. Microsoft windows 10 allowed writing comment or suggesting thoughts in Urdu. Including the speakers of Urdu language elsewhere in the world, there is lot of Urdu language data to be processed. Currently, most of the tools and approaches exist for English language (and other languages) that cannot be adopted for Urdu language [3] due to their script, morphological, and grammatical differences. As a result, Urdu language should be studied as an independent problem domain. While many have worked on different research problems for Urdu language already, there exist only very few works [24] [25] [26] for temporal entity extractions in Urdu language.

II. TEMPORAL ENTITY EXTRACTION

Temporal information extraction plays a crucial role in improved information access, in particular for creating timelines and detailed question answering. Temporal representation and reasoning in Natural Language (NL) is a nontrivial task due to: (1) the diversity of time expressions; (2) the complexity of determining temporal relations among events; (3) the difficulty of handling temporal granularity; and (4) other major problems in computational NLP (e.g., ambiguity, anaphora, ellipsis, and conjunction).

A. Importance of Temporal Entity Extraction

Extracting temporal information is not only important but accuracy and truthiness of extracted information does matter. Name entities provide the supporting hand to ensure the correctness of retrieved data. Person and location provide the designator of data while temporal entity helps to measure the accuracy and truthiness of data for specific period of time. In the following sub-section, we summarized that how extracted temporal entities' information can be useful for different NLP tasks.

B. Identification of Correct Time

Searching for Prime Minister of Pakistan may retrieve all PM's while expected data is about current PM. It is necessary to scroll down and go through the whole page by gazing contents to find required information. It wastes the time, may bore the user to scroll and read all paragraphs and as result user will leave the website and develop negative perception.

C. Prediction of Upcomming Events

Let us suppose an example text written both in English as well as Urdu.

" پاکستان کی بنیاداسلام ہے۔ یہ 14اگست 1947ء میں دنیا کے نقشہ پرابھرا۔ اسلامی دنیا کی تاریخ میں واحد ایٹمی طاقت ہے۔ اس کے زمینی خدوخال، ماحول اور ثقافت دنیا میں منفرد ہیں۔ اس کا 72 یومِ آذادی14 اگست 2019 میں منا یا جائے گا۔ " (Pakistan ki bunyad Islam hay. Yeh 14 August 1947 mein dunya kay naqshy per ubhra. Islami dunya ki tarekh mein wahid aetmi taqat hay. Iss key zameeni khadokhal, mahol, aur saqafat dunya mein munfrid hein. Iss ka 72 youm-e-azadi 14 August 2019 mein mayaya jay ga).

"Pakistan is based on Islam. It emerged on the map of world on 14 August 1947. It's only Atomic Power in the history of Islamic world. Its Land structure, environment and culture are unique in the world. Its 72 Independence Day will be celebrated on 14 August 2019. "

Upcoming event can be extracted and predicted by temporal entities i.e. 14 August 2019 in above paragraphs (written in Urdu language) represents the Independence Day of Pakistan. Extracting date from sentences or paragraph can be used to automatically calculate the remaining days of upcoming events (in this case) Pakistan's Independence Day. People can be easily facilitated, and security measurement taken.

D. Categorization of Real-Time and Retrospective Event

Classifying event as retrospective and real time from massive amount of data on social networks is challenging task. Temporal information plays vital role to resolve above mentioned challenge and also necessary to construct eventual timeline [5].

E. Evaluation Over Time

Temporal data can also be used to evaluate the popularity of personality, product, idea, music, games, celebrities and research topic being discussed among people over the passage of time. Time entity initiatively proposed in Natural Language was based on Algebra [6]. Acknowledging the importance of temporal entity extraction, temporal entity extraction as a task was highlighted in Message Understanding Conference (MUC-6) [7]. 'Date' is important subcategory of temporal entity [8] which plays vital role to extract exact, accurate and quick information from immense collection of unstructured and heterogeneous textual data. Questioning and Answering System, Text summarization, visualization, timeline, and semantic web development relay on temporal information [9].

Our work focuses on neglected cursive language Urdu to extract temporal entity 'date'. Urdu language in sub-continent: spoken, written and understood by more than 100 million people [10]. Urdu is national language of Pakistan and is understood by 75% of the population [11]. It is a mixture of different languages i.e. Turkish, Arabic, Persian, Hindi and Sanskrit [11]. Consolidation of different languages in Urdu language created writing variance and made Urdu language uniquely distinguishable then other languages [12]. The Govt. of Pakistan declared "dd\mm\yyyy "as official standard of date format for letters, reports, memos and other documents. Urdu dates can be classified into four major types i.e. Fully Qualified, Partially Fully Qualified, Deictic and Anaphoric.

In this research paper, we report our preliminary experiments for Urdu dates extractions for dates of all types using pattern matching techniques.

III. GRAMMATICAL STRUCTURE OF URDU LANGUAGE

A language allows penmanship by joining different letters together. Urdu [13] Arabic and Persian are the most popular cursive languages in sub-continent. Urdu language writing format made it distinguishable in context of characteristics. Nastalique writing style widely used that is diagonal and complex in nature by starting form right to left direction [14]. Subject Object Verb (SOV) is structural sequence of any Urdu language sentences. For example, the sentence "احمد نے پودوں کو پانی دیا۔" (Ahmad watered the plants) "follows the SOV format.

IV. TYPES OF TEMPORAL ENTITY

Categorically, temporal entity date can be classified into three types i.e. fully qualified date, deictic date and anaphoric date [15] [16].

A. Fully Qualified

A temporal expression [16] that consists of complete date information such as day, month and year i.e. dd/mm/yyyy بیس اکتوبر دوہزار اٹھارہ (20/10/2018).

B. Deictic

A temporal expression represents such type of date that required to further analysis. An expression of words required utterance of words [16]. For example: (1) (آج "Aaj"- 'today'), (2) (کل" kall" -'tomorrow') etc. A comprehensive but limited collection of deictic words used in Urdu language to represent time is given in Table 1.

TABLE I. DEICTIC WORDS

Deictic Words Representing Time			
فرصت	جمعرات	رات	لمحہ
مہلت	جمعہ	روز	دقیقہ
وقفہ	ہفتہ	یوم	ساعت
دورانیہ	اتوار	وار	پل
آغاز	تب	شب	گھڑی
شروع	کب	صبح	لحظہ
اختتام	ابھی	مہینہ	سیکنڈ
رُت	دیر	سال	اَن
تاریخ	تاخیر	برس	دم
حیات	جلدی	صدی	عہد

زندگی	اثنا	سحر	دور
باری	سردی	فجر	زمانہ
موسم	گرمی	دوپہر	آن
موقع	خزاں	سہ پہر	وقت
عمر	بہار	شام	قرن
دراز	اوقات	تڑکا	مدت
روزانہ	میعاد	سویرا	زمانہ
آج	ازل	سوموار	منٹ
کل	ابد	منگل	گھنٹا
پرسوں	عرصہ	بدھ	دن

C. *Anaphoric*

A case of deictic expression for which utterance of time varies according to the temporal expression as previously mentioned in the text [16]. For example: اُس سال (‘that year’), پچھلے ہفتے(‘last week’) and دو ماہ (‘two months’) etc.

V. TYPES OF TEMPORAL (DATES) ENTITY IN URDU LANGAUGE

Exhaustive analysis of Dates written in Urdu language in textual format depicted that it can be divided into four types i.e. Urdu Fully Qualified, Urdu Partially Fully Qualified, Urdu Deictic and Urdu Anaphoric. The detail description of all these types is given in proceeding section of paper.

A. *Fully Qualified Date*

A Fully Qualified Date (FQD) is “A Temporal Entity which gives detail information about event, action and act. It contains Day, Month and year. A date can be written in different format; depending on the langue i.e. English [16] has standard format yyyy/mm/dd while Urdu follows the dd/mm/yyyy format. For example, 25/07/20018 and ۲۵-پچیس جولائی-دوہزار آٹھارہ, respectively.

B. *Urdu Fully Qualified Date*

A date written in Urdu language which consists of Day, Month and Year is called Urdu Fully Qualified Date”. For Example,

(1) آٹھ فروری انیس سو اکانوے

Day, Month, Year and century can be represented in following manner:

- Roman Numbers 0,1, 2…,9 i.e. 02-10-2012
- Arabic Numbers i.e. (۰۳/۱۱/۱۹۹۱)
- Urdu words i.e. ایک، دو ،تین،چار---- چودہ اگست انیس دسمبر دو ہزار اٹھارہ
- Mix up of all i.e. 25-جولائی-2018

C. *Differetn Types of UFQD Regarding Processing*

Analysis showed that Fully Qualified Date (FQD) in Urdu language can be represented in different format so for convenient of understanding we suggested a name i.e. Hybrid Urdu Fully Qualified Date (HUFQD). These dates are given here:

- Numeric Day and Urdu Month/Year i.e. 25دسمبر 25مارچ انیس سو چالیس,دوہزارسات
- Urdu Day/Year and Numeric Month i.e. چار5 دوہزار دس8دوہزار تیرہ,بارہ
- Urdu D/Month and Numeric year i.e. 2008 مارچ یکم پندرہ جون 2009,

D. *Urdu Partially Fully Qualified Date*

A type of Date written in Urdu textual language which is missing one of the given i.e. Day, Month or year. For example, 26/2008, 08/2016 or 26/08 in English while in Urdu دس جولائی (07/2007). جولائی دوہزارسات (07/10),

E. *Deictic Urdu Date*

A type of date which cannot directly mapped to standard date format. It requires further analysis of context to give the purposeful meanings i.e. وقت، دن اب،تب، رات اور صبح وغیرہ

F. *Anaphoric Urdu Date*

A special case of Deictic date which requires utterance time which vary from time to time to conclude meaningful information i.e. اگلے سال، پچھلے دن، کئی سال.

VI. RELATED WORK

A considerable volume of research work exists for non-cursive languages especially for English, French, German, Dutch and Spanish [7] which achieved noticeable accuracy for developing mature artificial intelligence applications.

Cursive languages i.e. Arabic, Persian, Urdu and Hindi [17] neglected by researchers. Only few numbers of cursive languages were known publicly, due to lack of interest, inconvenience in processing, and unavailability of resources i.e. Lexicon, Databases, Dictionaries, Annotations schemes and Datasets [18]. To develop generic NLP applications, it is demand of time to include cursive language into research stream. Qaunzhi, li. et al. [5] used temporal module to filter out cluster of retrospective (old) event and real-time (new) events. Temporal information is crucial in differentiating between latest events. An approach developed [19] to automatically assign document event-time by extracting temporal expression from text. It helps to retrieve related document based on temporal values and finding relationship between them. Tianyong ho et al. [20] design a novel method TEER to extract and normalize temporal expression from heterogeneous clinical text. They use heuristic rules, summarization and automatic patterns learning. Developed system evaluated on two dataset i.e. English and Chinese clinical text which consists of 400 English and 1459 Chinese discharge summaries. Precision and recall for English and Chinese languages are 0.948, 0.877 and 0.941, 0.932 respectively. A sequencer system developed for analysis of temporal entities [21] existing in news articles and user generated unstructured contents. It is based on crawling, clustering, extracting and visualizing. Many annotation schemes i.e. PoS Tagging, Partial Parsing, Semantic Interpretation, case frame instantiation and discourse analysis were used to extract temporal expression from textual data [22]. A system ManTime developed [16] to explore the temporal expression using CRF. It used WordNet based feature but degraded overall temporal entity identification performance. Po-Yao Huang et. al [23] developed a system to monitor temporal event from social media. The main purpose of system is to monitor the temporal event on social media.

In Urdu literature, there is no proper exhaustive research work exist for temporal entities. In 2008 International Joint Conference on Natural Language Processing IJCNLP (IJCNLP) proposed a set of 12 named entities for South-Asian language including Temporal Entity i.e. Date and Time as single Entity [24]. A rule-based approach adopted in [25], which focused on date and time tags. They used Regular Expression (RE) to extract specific pattern of date i.e. 01.08.2015 or 01/01/2014. The same system also able to identify date like May 01, 2018 and achieved 90.83% F1-Measure. In our best knowledge, there is no detail discussion about different types and format of date in Urdu language. Lack of resources i.e. lexicon, gazetteers and dataset are the main factors to adopt rule base approaches. In [26] a generic name entity recognitions system used rules-based approach to extract name entities including Date from Urdu language which achieved considerable F1-Measure for specific pattern i.e. '1996' but unfortunately there exists no detail about types and format of dates in Urdu language. Another system developed to extract fluent information that is valuable for certain period. The claimed that many proposed systems focused on static information while mostly newswire text and Wikipedia are predominant temporal expression [2] precision and recall of Temporal Information Extraction were 0.50 to 0.99. In general, Temporal Expression identification performed by machine learning approaches based on lexical and morphological features [15]. Support vector Machine and Condition Random Field CRF give considerable results for Non-cursive languages respectively [27][28].

Central Language of Engineering (CLE) is working for Urdu language which offered different datasets available on website at affordable price. A Part of Speech Tagger (PoS) also developed by CLE and providing services online. It tags 100 words per attempt free. For further processing full access can be provided on request. A system developed by Central Language of Engineering (CLE) does not evaluate the Temporal Entities (TE). At the same website a small WordNet which contains data in UTF-8 format is also publicly available with some charges. Now, from last few decades' cursive languages being popular and attracted researchers to explore for development of NLP applications. Temporal Data in Urdu language introduced at very basic level internationally and nationally in different research papers but still no significant work proceeded in favor of Urdu Temporal Entity 'Date extraction'. In our best knowledge we are first one working on Temporal Entity 'Date' in Cursive Langue Urdu. A dataset developed by [29] publicly available for experimental purpose which is generic Name Entity Extraction. It consists of 12 different Name Entities including Date. It consists of 206 date tags with different written pattern/format of date. Existing Methods for Name Entity extraction for Temporal Entity extraction from Urdu language showed disappointing results which emphasized to explore Urdu language and introduce new methods and approach to develop NLP applications.

VII. PROCESSING ISSUES IN URDU LANGUGAE

Cursive langue Urdu is national language of Pakistan consists of two types of alphabets i.e. Joined and Non-Joined depending on the position of letter being used. For example, if ا، د،ر،ڑ،ز،ژ،ڈ،ز are used in the start of word these letters behave Non-Joined letter. For example (1) د لیرعلی (Daler Ali) here the letter 'د' is used in the beginning of word and the same words at the end considered as joined letter. In example (2) جدوجہد (Jad-o-Jahed) (struggle) the letter 'د' is joined. Urdu has different writing style, complex structure of letters and rich collection of alphabets. Some generic and basic issues regarding the Urdu language processing are [25]:

- No capital letters,
- Complex writing format,
- Words ambiguity,
- Right to left writing style,
- Improper sequence of words.

VIII. CHALLENGES IN TEMPORAL ENTITY EXTRACTION FROM URDU LANGUGAE SCRIPT

Different writing patterns in Urdu language created some issues for extracting dates. Many words can be used to represent date because Urdu is rich morphological language mix-up of many other languages i.e. Turkish, Arabic, Persian, and Hindi [30]. A dataset developed by [29] consists of 18 different formats of Urdu Date which tagged as <DATE>. Such patterns create hurdle to develop generic patterns-based rules.

A. Varying Pattern in Urdu Fully Qualified Date

In Urdu language Fully Qualified Date can be written in different format as mention in above section 5 which created processing issues. Every format is considered as individual pattern of date.

B. Deictic Date

In case of deictic date, determining semantically meaning of deictic word is serious issue. Duel meaning of words is also big challenge for Deictic Dates. For Example, a word گھڑی (Watch) can be a thing or time span. Some semantically ambiguous examples are given in Table2.

TABLE II. DEICTIC WORDS HAVING DUAL MEANINGS

Word	Examples
سویرا	سویرا بہت سمجھدار لڑکی ہے۔
	دن کا سویرا اور رات کا اندھیر۔
مارچ	مارچ میں بہار عروج پر ہوتی ہے۔
	سکول میں بچوں نے لانگ مارچ کیا۔
عمر	علی اور عمر دوست ہیں۔
	مد یحہ آپ کی عمر کتنی ہے؟
گھڑی	دیوار پر گھڑی لگی ہوئی ہے۔
	موت کی گھڑی مقرر ہے۔
اوقات	آپ سے ملاقات کے اوقات کیا ہیں؟
	انسان کو اپنی اوقات نہیں بھولنی چاہیے۔
جولائی	آمنہ جولائی ہے مجھے بھی دو۔
	علی اور اسلم جولائی میں گھومنے جاتے ہیں۔
حیات	ت نے گھریلو مسائل کی وجہ سے تعلیم ترک کر دی۔
	علامہ اقبالرح نے اپنی حیات میں شاعری کو طاقت بنائی۔
فرصت	پڑھائی سے فرصت ملے تو زندگی جیوں۔

	میری بیرون ملک تعلیم حاصل کرنے کی فرصت نہیں۔
پل	اگلے ہی پل وہ زندگی سے چلاگیا۔
	پچھلے سال پل ٹوٹ جانے سے کئی طالبہ ڈوب گئے۔
شام	دہشت گردوں نے شام کو ٹھکانہ بنایا ہوا ہے۔
	شام ہوتے ہی بجھا دیتا ہوں چراغ کو تیری یاد میں جلنے کو یہ دل ہی کافی ہے۔
دراز	علی اور سلویٰ عرصہ دراز سے دوست ہیں۔
	وہ دراز قد کا مالک ہے۔
	پستول درازمیں رکھی ہے۔
سحر	سحر کو خاموشی ہوتی ہے۔
	سحر کی کتاب میرے پاس ہے۔
روز	وہ ہر روز مجھے تنگ کرتا ہے۔
	مسلمان روزہ رکھتے ہیں۔
نور	نور کا تڑکا نکلتے ہی کسان کھیتوں کو چلے جاتے ہیں۔
	سمینہ اور نور بہنیں ہیں۔
تڑکا	تڑکا نکلتے سب کام پر چلے گئے۔
	او! بھائی سالن کو ذرا تڑکا لگا کے لانو۔

C. *Anaphoric Date*

Expressing meaning of Anaphoric date for a computer is very complex task as compared to human being. We human can determine the meaning of word by context and resolve the dual meaning problem easily but in case of computer it is very complex task. Although many solutions exist for Anaphoric and Deictic Date to extract from textual data but semantically determine the value of these Temporal Entities is very tough task which requires contextual information.

IX. SOLUTIONS

A. *Urdu Fully Qualified and Hybrid Urdu Fully Qualified Date*

A writing standard for date should be followed i.e. there must be space between day, month and year. Although Urdu language allowed writing date in different format but converting those date into standard format i.e. inserting space can be helpful for date extraction.

B. *Urdu Deictic and Anaphoric Date*

The words representing deictic date can be semantically understood as date with the help of Document Creation Time (DCT). For example, the word آج (today) can be converted into date by accessing the Document Creation Time. Similarly, the کل (tomorrow) and اگلے دن (next day) can also convert to standardize Fully Qualified Date.

X. METHODOLOGY

Extracting entities from textual data is performed by using different approaches depending on the availability of resources. In general, there are three approaches i.e. Rule-based, Statistic Base (Machine Learning) [25] and Hybrid [29]. Rules based approach required the deep knowledge of target language i.e. grammar, morphological and lexical insights. Rules are design based on patterns to extract specific entity [25].

Machine Learning is another approach, in which statistical information of documents is used to extract the entities from textual data. Such type of approach suitable when a large volume of resources i.e. Dataset, annotation schemes, WordNets and Part of Speech tagger exist. CRF, HMM, MaxEnt and Decision Tree are the common Models of Machine Learning [25]. Hybrid approach a mix-up of rule-based and statistic-based. It extracts the feature using rules and process the dataset using statistical models [29].

Existing Methods developed to extract temporal information from English and other non-cursive languages showed disappointing results which compel us to explore Urdu language as a research area to introduce new methods and approach to develop NLP applications. We decided to go ahead using Rule based approach i.e. Regular Expression (RE) for Temporal Entity Extraction. We developed generic Regular Expression which are evaluated on Urdu Dataset [29] which consists of Name Entities including Temporal Entities.

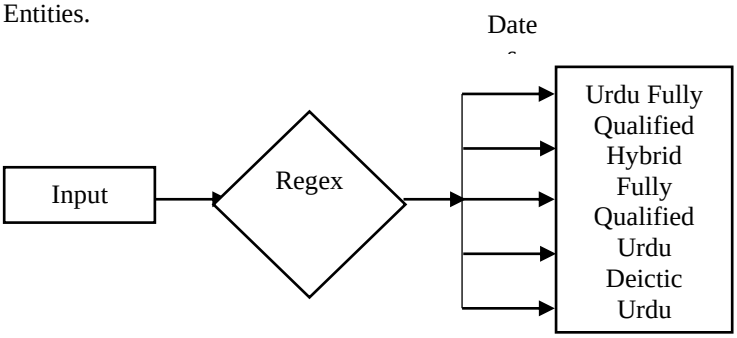

Figure 1. Methodology of Temporal Entity Extraction

XI. EXPERIMENT

We started our experiments on plain textual Urdu corpus [26] by neglecting the annotation tags. Complex structure and varying format of Urdu temporal entities converges our strength to use regular expressions for temporal entity extraction. In our exhaustively analysis we found different written pattern of Urdu Date i.e. Urdu Fully Qualified date, Urdu Deictic Date and Urdu Anaphoric Date.

A. *Dataset*

Dataset for Urdu language generally exists for name entity extraction with small number of instances which are:

- Enabling Minority Language Engineering (EMILLE) (only 200000 tokens) [31].
- Becker-Riaz corpus (only 50000 tokens) [32].
- International Joint Conference on Natural Language Processing (IJCNLP) workshop corpus (only 58252 tokens)
- Computing Research Laboratory (CRL) annotated corpus (only 55,000 tokens are publicly available data corpora [33].

In our knowledge there is no specific data set available for temporal entities extraction from Urdu language. We selected a dataset develop for name entity extraction [26]. It consists of 206 date tags including single month name, year or both of it. It is about National, Sports and International News including Urdu Fully Qualified, Urdu Hybrid Fully Qualified, Urdu Deictic and Urdu Anaphoric. Exhaustive analysis revealed that there are only 5-10 Fully Qualified Dates which made us impassive. It also revealed that 18

different date patterns are lying in limited date tags which created problem to write a generic regular expression for date extraction. The issue is resolved by writing generic regex and specific regex to extract temporal entities.

We decided to extend the existing data set by adding 200 Urdu Fully Qualified dates and 50 Urdu deictic words. The dataset extension detail is that we added 50 dates for UFQD and 150 dates for HUFQD.

Similarly, 50 deictic words were added 25 of them representing dates while 25 deictic words representing name entities. We placed these dates at different location in documents i.e. at sentence level, at the beginning, middle, and end of sentence. For example, پاکسان کی تاریخ میں چھ ستمبر دوہزار اٹھارہ کو سنہری الفاظ میں لکھا جائے گا۔ (Pakistan ki tareekh mein chey stmber do hazar atharah ko sunhari alfaz mein likha jay ga) in the sentence چھ ستمبر دوہزار اٹھارہ (chey stmber do hazar atharah) represents date which is placed in the middle of sentence.

B. Preporcessing

Some preprocessing measures are taken to prepare dataset:

- Fully Qualified Dates added manually in pre-existing dataset that having the same writing format standard i.e. spacing pattern, day, month year is separated by space.
- No, annotation tag used for dates, data added as plain text i.e. آٹھ اپریل دوہزار دو

C. Results Evaluation Parametrs

- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$
- $F\beta 1 = 2 * \frac{Precision * Recall}{Precision + Recall}$

XII. RESULTS AND DISCUSSION

All the results (Precision, Recall and F1-Measure) obtained by our approach are presented in tabular form while the accuracy is presented in pictorial form.

TABLE III. ALL TYPE OF DATES ON ORIGINAL DATASET

Type of Date	Pr.	Re.	F1
Numeric Year	0.91	1.00	0.95
Urdu Month and Year	0.58	1.00	0.77
Urdu Year	1.00	1.00	1.00
Urdu Month and Numeric Year	1.00	1.00	1.00
Numeric Day and Urdu Month	0.95	1.00	0.97
Only Urdu Month	0.18	1.00	0.30
Urdu Day and Month	0.50	1.00	0.67
UFQ Date and Urdu Hybrid FQ Date	0.95	0.95	0.95
Deictic and Anaphoric	1.00	1.00	1.00

TABLE IV. UFQD &UPFQD ON EXTENDED DATASET

Example	Date Type	Pre.	Re.	F1
2019جون	Numeric Year	0.96	0.92	0.94
8پانچ	Numeric Month	1.00	1.00	1.00
8 فروری	Numeric Day	0.94	1.00	0.97
تین اگست دو ہزار انیس	Urdu FQ Date	1.00	1.00	1.00
	Average	0.97	0.98	0.98

TABLE V. UFQD &UPFQD ON EXTENDED DATASET

Deictic date			
	Pre.	***Re.***	***F1***
Recognition	0.50	1.00	0.66
Retrieval	1.00	1.00	1.00

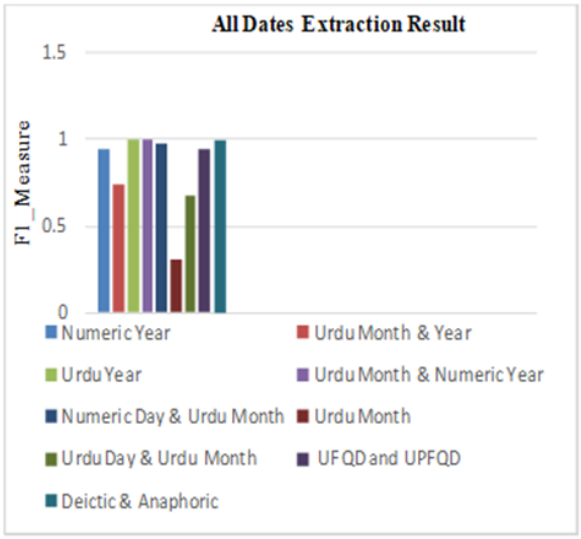

Figure 2. F1_ Measure on all types of date

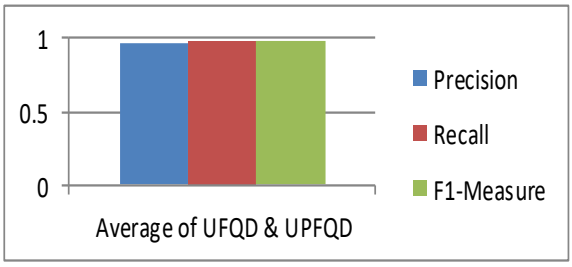

Figure 3. UFQD & UPFQD

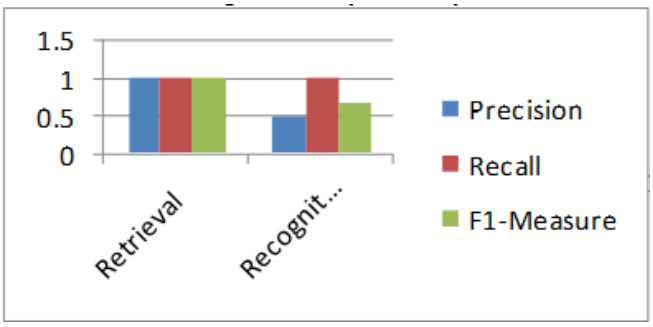

Figure 4. Deictic Date Analysis

XIII. CONCLUSION

Our approach comparing with existing temporal extraction work shows highly improved efficiency and accuracy, tackling different Urdu entity types using publicly available Urdu data corpus for our experiments, which was extended by our work. Also, a new date type for Urdu language was introduced. Some challenges and issues of other date types in Urdu textual Language i.e. Deictic and Anaphoric are also discussed in detail. This work can be extended and applied to the languages like Urdu, for example, Punjabi and Sindhi.

REFERENCES

[1] Al-Dyani, Wafa Zubair, Adnan Hussein Yahya, and Farzana Kabir Ahmad. "Challenges of event detection from social media streams." International Journal of Engineering & Technology 7, no. 2.15 (2018): 72-75.

[2] Ling, X., & Weld, D. S. (2010, July). Temporal Information Extraction. In AAAI (Vol. 10, pp. 1385-1390).

[3] Riaz, Kashif. "Concept search in Urdu." In Proceedings of the 2nd PhD workshop on Information and Knowledge Management, pp. 33-40. ACM, 2008.

[4] Cardoso, Nuno Francisco Pereira Freire, and Mário Jorge Costa Gaspar da Silva. "Semantic-flavored Query Reformulation for Geographic Information Retrieval."

[5] Li, Quanzhi, Armineh Nourbakhsh, Sameena Shah, and Xiaomo Liu. "Real-time novel event detection from social media." In 2017 IEEE 33rd International Conference on Data Engineering (ICDE), pp. 1129-1139. IEEE, 2017.

[6] Allen, James F. "Maintaining knowledge about temporal intervals." Communications of the ACM 26, no. 11 (1983): 832-843.

[7] Nadeau, David, and Satoshi Sekine. "A survey of named entity recognition and classification." Lingvisticae Investigationes 30, no. 1 (2007): 3-26.

[8] Ferro, Lisa, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. "TIDES 2005 standard for the annotation of temporal expressions." (2005).

[9] Mani, Inderjeet, James Pustejovsky, and Beth Sundheim. "Introduction to the special issue on temporal information processing." ACM Transactions on Asian Language Information Processing (TALIP) 3, no. 1 (2004): 1-10.

[10] Naz, Mamoona, and Sarmad Hussain. "Binarization and its evaluation for Urdu Nastalique document images." In INMIC, pp. 213-218. IEEE, 2013.

[11] Ghulam, Saqib Muhammad, and Tariq Rahim Soomro. "Twitter and Urdu." In 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1-6. IEEE, 2018.

[12] Kausar, Robina, and Muhammad Sarwar. "The History of the Urdu Language Together with Its Origin and Geographic Distribution." (2015).

[13] Shah, Zahra A. "Ligature based optical character recognition of Urdu-Nastaleeq font." In International Multi Topic Conference, 2002. Abstracts. INMIC 2002., pp. 25-25. IEEE, 2002.

[14] Hussain, Sarmad. "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, 2004.

[15] Ahn, David, Sisay Fissaha Adafre, and Maarten De Rijke. "Towards task-based temporal extraction and recognition." In Dagstuhl Seminar Proceedings. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2005.

[16] Filannino, Michele, and Goran Nenadic. "Temporal expression extraction with extensive feature type selection and a posteriori label adjustment." Data & Knowledge Engineering100 (2015): 19-33.

[17] Malik, Muhammad Kamran, and Syed Mansoor Sarwar. "Named entity recognition system for postpositional languages: urdu as a case study." International Journal of Advanced Computer Science and Applications 7, no. 10 (2016): 141-147.

[18] Riaz, Kashif. "Concept search in Urdu." In Proceedings of the 2nd PhD workshop on Information and Knowledge Management, pp. 33-40. ACM, 2008.

[19] Llidó, Dolores, R. Berlanga, and Mariá J. Aramburu. "Extracting temporal references to assign document event-time periods." In International Conference on Database and Expert Systems Applications, pp. 62-71. Springer, Berlin, Heidelberg, 2001.

[20] Hao, Tianyong, Xiaoyi Pan, Zhiying Gu, Yingying Qu, and Heng Weng. "A pattern learning-based method for temporal expression extraction and normalization from multi-lingual heterogeneous clinical texts." BMC medical informatics and decision making 18, no. 1 (2018): 22.

[21] Walenz, Brett, Robin Gandhi, William Mahoney, and Quiming Zhu. "Exploring Social Contexts along the Time Dimension: Temporal Analysis of Named Entities." In 2010 IEEE Second International Conference on Social Computing, pp. 508-512. IEEE, 2010.

[22] Woodward, Daryl. "Extraction and Visualization of Temporal Information and Related Named Entities from Wikipedia." (2001): 1-8.

[23] Huang, Po-Yao, Junwei Liang, Jean-Baptiste Lamare, and Alexander G. Hauptmann. "Multimodal Filtering of Social Media for Temporal Monitoring and Event Analysis." In Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, pp. 450-457. ACM, 2018.

[24] Liao, Wenhui, and Sriharsha Veeramachaneni. "A simple semi-supervised algorithm for named entity recognition." In Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, pp. 58-65. Association for Computational Linguistics, 2009..

[25] Singh, U., Goyal, V., &Lehal, G. S. (2012). Named entity recognition system for Urdu. Proceedings of COLING 2012, 2507-2518.

[26] Riaz, Kashif. "Rule-based named entity recognition in Urdu." In Proceedings of the 2010 named entities workshop, pp. 126-135. Association for Computational Linguistics, 2010.

[27] Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. "A training algorithm for optimal margin classifiers." In Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152. ACM, 1992.

[28] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001).

[29] Khana, Wahab, Ali Daudb, Jamal A. Nasira, and Tehmina Amjada. "Named entity dataset for urdu named entity recognition task." Organization 48 (2016): 282.

[30] Hardie, Andrew. "Developing a tagset for automated part-of-speech tagging in Urdu." In Corpus Linguistics 2003. 2003.

[31] Baker, Paul, Andrew Hardie, Tony McEnery, and B. D. Jayaram. "Corpus data for South Asian language processing." In Proceedings of the 10th Annual Workshop for South Asian Language Processing, EACL. 2003.

[32] Becker, Dara, and Kashif Riaz. "A study in urdu corpus construction." In Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12, pp. 1-5. Association for Computational Linguistics, 2002.

[33] Kanwal, Safia, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. "Urdu Named Entity Recognition: Corpus Generation and Deep Learning Applications." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 19, no. 1 (2019): 8.

[34] Andrew Borthwick, John Sterling, Eugene Agichtein and Ralph Grishman “Description of the MENE Named Entity System as Used in MUC-7” http://acl.ldc.upenn.edu/muc7/M980018.pdf Accessed on December 2011.