Quantifying the Use of English Words in Urdu News-Stories

Mehtab Alam Syed[1], Arif Ur Rahman[2], Muzammil Khan[3]
[1]Department of Computer Science, Free University of Bozen-Bolzano, Bolzano, Italy
[2]Department of Computer Science, Bahria University, Islamabad, Pakistan
[3]Department of Computer Science, University of Swat, Swat, Pakistan
msyed@unibz.it, badwanpk@bui.edu.pk, muzammilkhan86@gmail.com

***Abstract*—The vocabulary of Urdu language is a mixture of many other languages including Farsi, Arabic and Sinskrit. Though, Urdu is the national language of Pakistan, English has the status of official language of Pakistan. The use of English words in spoken Urdu as well as documents written in Urdu is increasing with the passage of time.**

The automatic detection of English words written using Urdu script in Urdu text is a complicated task. This may require the use of advanced machine/deep learning techniques. However, the lack of initial work for developing a fully automatic system makes it a more challenging task. The current paper presents the result of an initial work which may lead to the development of an approach which may detect any English word written Urdu text. First, an approach is developed to preserve Urdu stories from online sources in a normalized format. Second, a dictionary of English words transliterated into Urdu was developed. The results show that there can be different categories of words in Urdu text including transliterated words, words originating from English and words having exactly similar pronunciation but different meaning.

***Keywords*-Urdu Text Processing, Urdu Transliteration to English**

I. INTRODUCTION

Urdu is the national language of Pakistan and is spoken in many countries. Estimates show that Urdu is spoken by 164 million people around the world [1], [2]. The word “Urdu” is basically taken from Turkish language and it means “army”. The vocabulary of Urdu language contains words from many languages including Persian, Arabic, Turkish, Sanskrit, and Hindi.

The issue of using Urdu as official language is raised in Pakistan every once in a while. Once even the supreme court of Pakistan ordered to use Urdu in official communications. However, the task of language switching is very challenging as people almost do not know the pure Urdu in daily life. People mostly use a mix of Urdu and English in daily life conversations. The reasons of not knowing the words is the use of a mix of Urdu-English language in daily life.

It is a challenging task to write words which are borrowed from other languages (loan words) with different alphabets and sound inventories. Typically, loan words are transliterated, i.e. replaced with approximate phonetic equivalents. Some approaches have already been developed for automatically transliterating Japanese to English, English to Japanese, Arabic to English and English to Arabic [3], [4]. Though, Google has developed a tool for English to Urdu transliteration, our analysis shows that it is still in its initial stages. Moreover, there are no tools available to automatically detect English words transliterated to Urdu in Urdu text. The tools developed as a part of the current work uses a lexicon based approach for detection of English words transliterated in Urdu. This can prove to be a starting point for developing approaches using more sophisticated techniques like neural networks e.g. Long Short Term Memory(LSTM) networks.

II. RELATED WORK

Digital preservation can be understood as ‘the ability to sustain the accessibility, understandable and usability of digital objects in the distant future regardless of changes in technologies and in the designated communities’ [5]. The data which is needed for long term plays an important role in our daily life. These data is to be preserved in an organized format so that it will be retrieved whenever it is need. The data which is preserved can be organized using time or other parameters which is important according to the perspective [6].

Long-term preservation approaches have comprised of emulation, migration, normalization, and metadata or some combination of these. Most existing work has focused on applying these approaches to digital objects of a singular media type: text, images, databases, video or audio. There is also a need to consider the preservation of composite, mixed-media digital objects which is a rapidly growing category of resources. It describes an integrated, flexible system that is to be developed, which leverages existing tools and services and assists organizations to dynamically discover the optimum preservation strategy as it is required. The system captures and periodically compares preservation metadata with software and format registries to determine those objects at risk. By making preservation software modules available as Web services and describing them semantically using a machine-process able ontology, the most appropriate preservation service for each object can then be dynamically discovered, composed and invoked by software agents. The growing array of available preservation tools and services can be integrated to provide a sustainable, collaborative solution to the long-term preservation of large-scale collections of complex digital objects [7].

There are other approaches available which proposed to look at, inspect, and report the stream of news data, its content, and information for four noteworthy daily papers

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

from generation and sourcing, through editing and printing, to dissemination to end clients [8]. It was the creator's trust that their report may pinpoint the high impact factors in the work process where libraries and other memory associations could catch basic news substance and metadata; and guarantee the long awaited survival and availability of the American journalistic record. The outcomes, however as anyone might expect, indicated little consistency among work processes. The results also indicated that a high level of standardization would be vital for fruitful preservation of repositories. There are other studies available which link news stories in the process of preservation [9]. The identification of transliterated words can also be helpful in linking news stories.

There are three unique techniques to capture online news and other web resources, namely by authoring system, by browser and by web crawler [10]. The web crawler approach is followed to preserved the digital news to build the archive. Web crawler is a program which analyze and extract desired information from the web page or pages.

The other aspect of the research after building the archive is to find the transliterated English words in Urdu news archive. The dictionary meaning of translation is the process of changing something that is written or spoken into another language, whereas transliteration is to write or describe words or letters using letters of a different alphabet or language [11].

Transliteration process is the replacement of words from source language to the approximate phonetic or spelling equivalents in the target language. Transliterating names between languages that use similar alphabets and sound systems are very simple, since the phrase mostly remains the same. The transliteration becomes difficult when transliterating between languages having different sound systems and having different writing system [12].

It is challenging to translate all the words or lexicons across languages with different sounds and phonetics. These words or lexicons are mostly transliterated from source language to the target language with its approximate sounds and phonetics. For example, "computer" in English comes out as "konpyuutaa" in Japanese [13]. Same phonetics words or lexicons used in different languages worth a lot because of it gives an interesting results about the different origin words are used across different languages.

A. Translation vs. Transliteration

Translation is the process of translated words or text from one language to another language. It gives the equivalent semantic meanings in the destination language in which it is to be converted. Translation is the spoken word or text meaning of a word in source language and translated in the destination language. e.g. To translate the English language words into Urdu language 'Word' to لفظ, 'Sentence' to جملہ respectively.

Transliteration is converting the text from one language to another and does not render meaning. For example the English word 'break' is very frequently used in Urdu and written as 'بریک', 'make' as 'میک', and 'cake' as 'کیک' respectively.

III. CORPUS BUILDING

Though there are some corpora available which could be used but they lack certain information and are not openly available. The purpose of building a new corpus is that the work could be easily taken forward after properly annotating the tokens. The annotating of tokens is a basic requirement for developing more sophisticated approaches i.e. using deep learning.

A. Workflow for Archive Building

The first step is to identify the sources which are used for building the archive of News stories. Three News sources are selected for building the archive which include popular online Urdu newspapers.

The first step is to analyze the 'HTML' source of each News source, identify the URLs of in 'href' tags which contain the News story. Identify the HTML tags which carry the information about the news story in each news story web page. Once analyzing the News web story is completed the data from the news is extracted and converted to the normalized format and thus preserved. This task is completed using a different extractor for each source which takes into account the specific structure of the HTML of the source. A normalized format is a standard format which can be accessed using standard tools and technologies. The story and its related information is further categorized into metadata and the story itself with hash comparison. The metadata is then categorized into 'explicit metadata' and 'implicit metadata'. Explicit metadata is directly available in with the story in HTML tags and includes information like title, publishDate, category, description, format, keywords, and language. Explicit metadata parameters are identified, extracted and mapped on Dublin core metadata standard. Implicit metadata on the other hand needs to be extracted from the title and story. The title and story are tokenize so that all the words in the story and title are separated and its frequency of occurrence is calculated and added in the implicit metadata respectively.

Individual words are referred to as Tokens (at least in languages like English) and tokenization is taking a text or set of text and breaking it up into its individual words. Tokenization is usually done for various tasks like finding frequencies of words. In Urdu language the tokenization is different from English as in English language the words are separated with a space in between them but in Urdu language two words may or may not need to be separated using a space character e.g. بڑا میز - are two words without a space between them. The tokenization for building the archive has some incorrect which have some error rate while finding the transliterated words in the archive.

Metadata is stored in `metadata.xml`, story is stored in `story.xml` and hash is calculated for comparison is stored in `hash.xml` respectively.

1) Hash Comparison and Issues: The hash is calculated for the story and metadata digital objects to identify if a previous version of the story is already stored in the archive. Hash is calculated for the story and metadata is compared with the `hash.xml` of all stories.

- The calculated hash should ignore the `storyId` as the storyId will be unique for all the stories.
- If the calculated metadata hash of the story is same with any story metadata hash in `hash.xml` and the calculated hash of story is different from story hash in `hash.xml` then a new version 'v1', 'v2' or so is updated in the archive.
- If the calculated story hash of the story is same with any story hash in `hash.xml` and the calculated hash of metadata is different from metadata hash in `hash.xml` then a new version 'v1', 'v2' or so is updated in the archive.

The complete workflow of building the archival of Urdu news stories is presented in figure 1.

The issue with Urdu language is that there is a very small number of sources from where digital-born text can be collected. Even the digital version of the most famous Urdu dictionary i.e. Ferozul Lughat (فیروز اللغات) is a scanned version of the printed book and does not allow to search words using latest searching techniques. The issue with the scanned versions of books is that Urdu OCR technology is still very weak and does not perform good in many situations [14], [15]. Each story is a combination of metadata, some associated files and the text of story.

The metadata is stored in a `metadata.xml` and includes story ID, author name, publication date & time, category, keywords and subject. All the metadata is taken automatically from for the source except the Story ID. The Story ID is composed of the publication date and time, a random number and version number e.g. `2016051600001347300507v0`. A default date and time i.e. `000101010000` is added in case the publishing date and time is not known. It is assigned to each story at the time of archival. Sometimes stories may be updated after publishing them. In such cases, a second version of the same story is archived. The categories include sports, showbiz, poetry, cooking and horoscope. The analysis of usage of words (English or Urdu) is done on the basis of category. The associated files are stored in a folder and may include images, audio and video included in a story. The `story.xml` file includes the text of story which is divided into paragraphs.

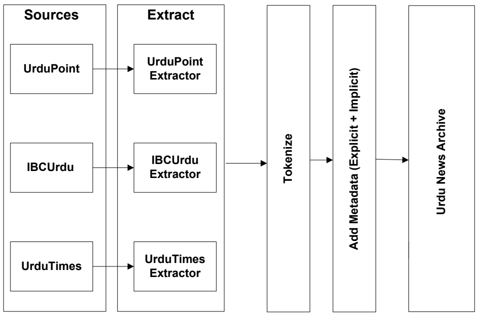

Figure 1: Workflow

IV. LEXICON BUILDING

A manual mechanism was used to build the lexicon which was further used for transliterated words in the Urdu news archive. A tool is created for adding the lexicon from the text which is provided as input It will show the lexicon one by one to manually add by user input as if the word is an Urdu word then it will be added as in `UrduDictionary.xls`, Similarly as if the word is a transliterated English words it will be added as in `UrduEnglishDictionary.xls`, On the other hand the ignored words are added in `otherDictionary.xls` respectively. More than **2100** transliterated English words are added in `UrduEnglishDictionary.xls` which is further used to process the build archive.

The corpus was used to build two lexicons i.e. an Urdu and English. This may seem a bit trivial but this is the first attempt of this sort and helped in identifying some issues which may need to be taken care of while developing an Urdu lexicon through an automated process.

The Urdu lexicon contains all valid Urdu words in the corpus. The following rules are followed.

- Nouns are included without considering language.
- Some words may be written in multiple form e.g. رحمان and رحمن (the same words written in two different forms).
- There are some words which are neither English words not Urdu words. However, these words are derived from English words and are frequently used in Urdu. Table I presents some examples of such words. These words are included in the Urdu lexicon.
- There are some words in Urdu and in English which has the same pronunciation like the words presented in table II. These words are added in the Urdu lexicon.

A second lexicon is developed i.e. English/Urdu lexicon which contains transliterated English words written in Urdu e.g. challenging and archive can be written as چیلنجنگ and آرکائیو respectively. The following rules are followed for building the lexicon.

TABLE I. URDU WORDS HAVING ENGLISH ORIGIN

English	Urdu	English	Urdu
scheme	اسکیمیں	agent	ایجنٹوں
agenda	ایجنڈے	agency	ایجنسیاں
community	کمیونٹیوں	ambulance	ایمبولینسوں

TABLE II. WORDS WITH SIMILAR PRONUNCIATION

English	Urdu	English	Urdu
eye	آئی	kiss	کِس
server	سرور	cone	کَون
peace	پیس	same	سیم

- English text written using English letters is ignored.
- English homophones e.g. lift and left, are written the same in Urdu as presented in table III. Therefore, one of the English words is listed in the English lexicon.

V. CORPUS ANALYSIS

The tool for building the corpus is completed and only in the past few days a corpus which contains more than six hundred stories was developed. The contains a total number of:

- Tokens in the corpus: 117393
- Unique tokens in the corpus: 10914
- Urdu Words in the corpus: 101147
- English Words in the corpus: 9962
- Unique Urdu Words in the corpus: 7770
- Unique English Words in the corpus: 1038

The table IV presents further statistics about the corpus. The size of the corpus will grow with the passage of time and these statistics may be affected by the change in size of the corpus. The last column in the table presents the number of ‘Other Words’ which may be unrecognized words. The unrecognized words include words which are not properly tokenized and words of languages other than English & Urdu.

A. Tokenization and its issues

Tokenization is the initial step for every language processing tasks e.g. machine translation, sentence ending detection, information retrieval and information extraction and preservation of corpus from which information is to be extracted for the long term which is further to be used by different standard tools. In most of the Asian languages the space is to be used after each character placement which is to be omitted to form the token as token consist of each different single word. Joiners are the characters which can occupy the initial, medial or final forms in the word.

TABLE III. WRITING HOMOPHONE IN URDU

Urdu	English	English
لِفٹ	lift	left
مَین	man	main
ھیٹ	hat	hate
کیمپس	hat	hate
کیریئر	career	carrier
نیس	niece	Nice (city)

TABLE IV. CATEGORY WISE STATISTICS

Category	A	B	C	D
Sports	243	2674	487	1004
Showbiz	143	1677	286	598
Cooking	100	487	66	154
Horoscope	38	288	17	22
Kids	36	1338	17	86
International	23	802	132	207
Women	18	399	27	23

A: Number of Stories
B: Number of Distinct Urdu Words
C: Number of Distinct English Words
D: Number of Distinct Other Words

Tokenization issues may be of various types. For example some English words cannot be separated from another Urdu Word which means the token being analyzed is basically a combination of an Urdu word and an English word like ڈیزائنرمیں, کرکوالیفائی , کیرئیرکا and اولمپئنزکیساتھ. Moreover, there may be tokens which are a combination of a word and a number. Additionally, some tokens may be a combination of an Urdu word and a word written using English letters.

Tokenization is a step required for applying other algorithms for accurate NER and POS tagging. However, the lack of an efficient tokenizer make it even difficult to process Urdu text for any purpose. The following issues make Urdu NER a challenging task. The examples of Urdu text ambiguities are presented in Table V.

1) **Punctuation Ambiguity:** Urdu language has punctuation ambiguity as punctuation marks are also used to specify the range, inside date, acronyms and abbreviations [16]. NER also used to extract the dates from the text, when it extract the date which separates the day, month and year with (-), it creates confusion because (-) is used for sentence termination. Therefore, smart techniques are required for accurate NER.
2) **Diacritics:** Some Urdu writers may use diacritics on some words which may change the word. For example حُسین is a noun but حسین is an adjective.
3) **Word Ending with Non-connectors:** Urdu characters are divided into two groups i.e. connector and non-connector. Whenever a word ends with a non-connector, there is no need to insert space, as native speaker can easily differentiate it from other word [2], [16]. For example قراردادِپاکستان, قرارداد and پاکستان are two words with no space between them. Space is not inserted between words as the word قرارداد ends with a non-connector which makes it difficult to extract پاکستان as location name. Therefore, such situations will require special techniques to separate the words for accurate NER.
4) **Sentence Boundary Ambiguity:** A native speaker can easily define the boundary of sentences by considering the context of words [2], [16]. It is complicated to identify the end of a sentence due to punctuation ambiguity and no capitalization. NER will fail to extract the relationship of the entities as well as the position of occurrence of the entities at sentence level if sentences are not segmented properly.
5) **Sentence Segmentation** The sentence terminator (-), in Urdu text is a confusing punctuation mark. Moreover, in Urdu text, a sentence is usually followed by dash (-) to terminate the sentence, which is not consistent. Most of the time, sentence also terminated with other word such as mark of exclamation (!), question mark (-), ellipsis (…) and bullets. Therefore, some techniques need to be developed for Urdu sentence segmentation. The technique to be

developed may be based on existing approaches for other languages.

6) **Order of Words:** Typically in Urdu, subject, object and verb sequence is used to make sentences but sometimes the order of words does not matter [16], [17]. The position of words can change the meaning of a sentence in Urdu. Therefore, it may be important to consider the whole context in which words are used to accurately identify named entities.
7) **Spaces Ambiguity:** Urdu language has agglutinative property which may lead to ambiguous spaces between words. Following are the scenarios in which space should be neglected.
8) **Word Segmentation:** Word Segmentation process needs to determine the boundaries of the words. In Urdu language the word is not separated by space so it is difficult to measure the boundaries of the word in Urdu language.
9) **Compound Words and Suffixation:** In Urdu words, two nouns, verbs and an adjective can be combined with or without the use of و, ا or اور to produce a single semantic word. Additionally, a suffix also generates words which may convert common nouns to proper nouns. If حسین و جمیل is considered a unit the meaning of the word is beautiful while حسین and جمیل are the names of the persons. The NER will consider حسین and جمیل two separate entities while it is a unit when و appears between these two words which is not an entity.
10) **Reduplication:** Reduplication of words is divided into two categories i.e echo and full, which may lead to white space between words. In echo reduplication one set of words is deviated where as in full reduplication a word is repeated twice. In a sentence فیصل •فیصل, فیصل مسجد جا ئے گا produces the reduplication of words while فیصل is the name of a person and فیصل مسجد is the name of the mosque. In this case NER could not be able to extract the entities properly.
11) **Loan Words:** Urdu language has borrowed many words from other languages. The loan words should be considered single semantic words.
12) **Nouns Ambiguity:** In Urdu language, nouns ambiguity may occur when proper nouns are used as common nouns. For example, identical other language words and person name. Also, most of the proper nouns have a white space between them, which should be neglected. NER will not be able to differentiate between common and proper nouns which may lead to false positive.

VI. Conclusions and Future Work

There are basically two motives for developing the corpus. First, to archive the news in a normalized format which allows easy access in the future. Second, to analyze the corpus for detection of English words transliterated in Urdu and get an idea of how many English words are used in Urdu. The size of the corpus is definitely a factor which may influence the frequency of words. Moreover, if a corpus contains news from a specific category then the set of frequent may also be different. However, the presented work is a step towards developing an automatic approach for developing a corpus and identifying English words in the corpus.

Further research is required to develop an approach which does not use dictionaries and can extract English words from a document using modern information retrieval and deep learning techniques. Moreover, the corpus will need to be properly annotated before it can be used to develop neural networks for the same task.

References

[1] M. Humayoun, H. Hammarström, and A. Ranta, "Urdu morphology, orthography and lexicon extraction," *Chalmers University of Technology, Master*, 2006.

[2] Z. Rehman, W. Anwar, and U. I. Bajwa, "Challenges in Urdu text tokenization and sentence boundary disambiguation," in *Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP 2011)*. Citeseer, 2011, p. 40.

[3] K. Knight and J. Graehl, "Machine transliteration," *Comput. Linguist.*, vol. 24, no. 4, pp. 599–612, Dec. 1998. [Online]. Available: http://dl.acm.org/citation.cfm?id=972764.972767

[4] N. Habash, A. Soudi, and T. Buckwalter, *On Arabic Transliteration*. Dordrecht: Springer Netherlands, 2007, pp. 15–22. [Online]. Available: http://dx.doi.org/10.1007/978-1-4020-6046-5_2

[5] S. Rabinovici-Cohen, R. Cummings, and S. Fineberg, "Self-contained information retention format for future semantic interoperability." in *SDA@ JCDL/TPDL*, 2014, pp. 4–15.

[6] Muzammil Khan and Arif Ur Rahman, "A Systematic Approach Towards Web Preservation," *Information Technology and Libraries*, vol. 38, no. 1, pp. 71–90, 2019.

[7] S. Ross, "Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries," *New Review of Information Networking*, vol. 17, no. 1, pp. 43–68, 2012.

[8] J. Alverson, K. Leetaru, V. McCargar, K. Ondracek, J. Simon, and B. Reilly, "Preserving news in the digital environment: Mapping the newspaper industry in transition," *crl. edu*, 2011.

[9] M. Khan, A. Ur Rahman, and M. Daud Awan, "Term-based approach for linking digital news stories," in *Digital Libraries and Multimedia Archives*, G. Serra and C. Tasso, Eds. Cham: Springer International Publishing, 2018, pp. 127–138.

TABLE V. URDU TEXT AMBIGUITIES

Punctuation Ambiguity	23-03-1940 کو قرارداد پاکستان منظور ہوئی
Order of Words	کھیلتے کھیلتے بچے میدان سے باھر چلے گئے
Compound words	مادر ملت , وزیر خارجه
Reduplication	چلتے چلتے , کبھی کبھی
Loan Words	چیک آؤٹ ,سمارٹ فون
Nouns	فیصل ,سرور
Bidirectional	علامه اقبال 9 نومبر کو پیدا ھوــئے 23-03-1940 کو قرارداد پاکستان منظور ہوئی

[10] S. Farrell, “A guide to web preservation,” *UKOLN / ULCC*, 2010.

[11] K. Regmi, J. Naidoo, and P. Pilkington, “Understanding the processes of translation and transliteration in qualitative research,” *International Journal of Qualitative Methods*, vol. 9, no. 1, pp. 16–26, 2010.

[12] Y. Al-Onaizan and K. Knight, “Machine transliteration of names in arabic text,” in *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, ser. SEMITIC ’02. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002, pp. 1–13. [Online]. Available: http://dx.doi.org/10.3115/1118637.1118642

[13] K. Knight and J. Graehl, “Machine transliteration,” *Comput. Linguist.*, vol. 24, no. 4, pp. 599–612, Dec. 1998. [Online]. Available: http://dl.acm.org/citation.cfm?id=972764.972767

[14] S. T. Javed, S. Hussain, A. Maqbool, S. Asloob, S. Jamil, and H. Moin, “Segmentation free nastalique Urdu OCR,” *World Academy of Science, Engineering and Technology*, vol. 46, pp. 456–461, 2010.

[15] N. Sabbour and F. Shafait, “A segmentation-free approach to Arabic and Urdu OCR,” in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2013, pp. 86 580N–86 580N.

[16] K. Riaz, “Rule based named entity recognition in Urdu,” in *Proceedings of the 2010 Named Entities Workshop*, 2010, pp. 126–135.

[17] S. Naz, A. I. Umar, S. H. Shirazi, S. A. Khan, I. Ahmed, and A. A. Khan, “Challenges of Urdu Named Entity Recognition: A Scarce Resourced Language,” *Research Journal of Applied Sciences, Engineering and Technology*, vol. 8, no. 10, pp. 1272–1278, 2014.