CLE Urdu Books N-grams

Farah Adeeba, Qurat-ul-Ain Akram, Hina Khalid, Sarmad Hussain
Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science,
University of Engineering and Technology, Lahore
firstname.lastname@kics.edu.pk, ainie.akram@kics.edu.pk

Abstract

The paper presents the development of first publically available Urdu N-grams extracted from different books. For the best representation of N-grams, large amount of Urdu corpus is collected from books covering different domains. The automatic cleaning of 37 million words corpus is discussed. The domain-wise N-grams are extracted which can be used in different Natural Language Processing and Information Retrieval applications.

1. Introduction

Reliable, well balanced and sizeable corpus is important for the development of mature Natural Language Processing (NLP) and Information Retrieval(IR) applications. These applications rely on language model which represents the characteristics of any language. N-gram is one of the most explored and used probabilistic language model to develop such applications. Normally, data sparsity issue appears if N-grams are computed from the corpus, which covers limited contextual information of words. Hence, large amount of words corpus is required which has rich contextual information of words, having a reasonable large number of N-grams with minimum data sparsity.

In addition, a balanced corpus is required, which covers reasonable domains for language coverage. In literature two widely used Urdu corpora [1,2] are reported. These corpora are extracted from Urdu magazine and news. CLE Urdu Digest [1] is publically available corpora varying from 100K[1] to 1M[2], that can be used for language modeling and N-gram extraction. These available Urdu corpora cover limited domains. In addition, the size of these corpora is also not too large. Therefore, new text corpus, having a reasonable domain and size is collected and presented in this paper. The corpus distribution into domains, automatic corpus cleaning and generation of N-grams are discussed in this paper.

2. Literature Review

A lot of effort has been carried out for the development of structured publically available text corpora in different European and Asian languages. Among them, a majority of the work focused on the development of corpus for English language [8,3]. The researchers use these publically available corpora to develop the language model for different applications. Effort is now focused on the development of N-grams and annotated N-grams, so that language model can be made available to users. Google Web IT 5-grams is publically available N-grams corpus collected from Google web books. The Version 2 of this corpus contains more than 8 million books [6]. This corpus includes N-grams (up to 5-gram) annotated with part-of-speech tags.

A large amount of text corpus for Arabic language is reported by Leipzig Corpora Collection [7]. This corpus contains Arabic text of different countries, including Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Lebanon, Mauritania, Morocco, Oman, Palestine, Qatar, Sudan, Syria, Tunisia, United Arab Emirates and Yemen. This corpus is crawled from online news websites of different countries. The country wise distribution of this corpus is also reported.

Persian language belongs to Arabic script and shares same complexities of Urdu including writing style and word segmentation issues. AleAhmad et al. [5] report a standard text corpus called Hamshahri for Persian language. This text corpus is crawled from online website of Hamshahri newspaper. The corpus is categorized into 82 different domains. This corpus contains a lexicon list of 417,335 words. Darrudi et al. [4] report character N-grams (up to 5 grams) computed on Hamshahri text corpus.

The effort has been carried out for the development of the Urdu text corpus. Ijaz and Hussain[2] report Urdu text corpus of 18 million words crawled from

[1] http://cle.org.pk/clestore/urdudigestcorpus100k.htm
2 http://cle.org.pk/clestore/urdudigestcorpus1M.htm

two news websites. This corpus is collected from different news domains, including finance, culture, science, etc. This corpus contains 104,341 unique words. This is not publically available due to licensing constraint. Urooj et al. [1] also report the 100K words Urdu corpus called of CLE Urdu Digest Corpus extracted from Urdu Digest magazine. The extracted data is categorized into 13 different domains. Later, the same approach is used to develop 1 million words Urdu Digest corpus. The reported 100K Urdu tagged corpus is also annotated with Part of Speech (POS) tags. All these Digest corpora are publically available. These corpora are not sufficient to develop the well representing language model for Urdu. Hence there is a need to collect the large Urdu text corpus covering a diversity of domains so that representative language model can be developed.

3. N-grams Development

For the development of N-grams, the first step is the acquisition of Urdu text corpus, which should cover a diversity of different domains. After acquisition, corpus is segregated into two main genres i.e. poetry and prose. Genre classification is done manually. Genre specific corpus is cleaned based on the Urdu characteristics and manual analysis of books content. After cleaning, the N-grams of the cleaned corpus are computed. The details of each process are given in sub-sequent sections.

3.1. Corpus Acquisition

To address the need for coverage of various domains of Urdu text, Urdu books are crawled from the web. A total of 1,399 books are collected from an online Urdu library [9]. The licensing information of these books is unspecified therefore N-grams of this corpus are reported and released publically under institutional license. These books are available in Unicode format. In first pass, each book is manually analyzed and categorized as poetry book or prose book. This classification is done by reading the content of books. The percentage of content is used for the categorization of book in specific domain. After this manual analysis and categorization, a total of 861 books having 37,680,293 words belong to prose category and 507 books having 309,486 words belong to poetry domain. During books distribution into different domains, there are 31 books which contain non-Urdu content therefore these books are not considered for domain classification.

3.2. Domain Wise Corpus Distribution

The books belong to prose are further analyzed to classify them into different domains. Therefore a complete manual pass has been carried out and prose books are categorized into 18 different domains, including articles, biography, character representation, culture, foreign literature, health, history, interviews, letters, magazines, novels, plays, religion, reviews, science, short stories, travel and Urdu literature. The domain wise books information is given in the Results section.

3.3. Corpus Cleaning

Although, the corpus is available in Unicode file format, but still there exists some web based content in books. Therefore books are further processed to remove such erroneous text such as HTML tags, URL, and Non-Unicode text. This raw corpus is processed to extract the words based on space tokenization. The analysis of this list shows that words are not properly space delimited in this corpus and a sub word or more than one words are resulted as single Urdu word after tokenization. Some of the erroneous words examples extracted from the raw corpus are listed in Table 1.

Table 1: Examples of erroneous words

Word	Frequency
توحید)	13
میں—	13
الصف:61	14
ادبِ	91
فضاءسے	1

This analysis shows that the extracted N-grams on this corpus will not give desired information and will contain erroneous contextual information of words.

Therefore, to address this issue, the corpus cleaning process needs to be done for proper space insertion between words before extraction of N-grams. To aid the cleaning process for Urdu text, a cleaning tool[3] is also available to assist the manual cleaning. The manual cleaning of this 37 million words is not feasible, therefore extracted word list is analyzed and semi-automatic corpus cleaning process is devised. After analysis of word list and corpus, following cleaning issues are extracted. The automatic way to address the issue is also discussed subsequent sections.

[3] http://www.cle.org.pk/software/langproc/corpuscleaningH.htm

3.3.1. Normalization. The Urdu words which can be written using different sequence of Unicode characters also exist in the corpus e.g. ۂ can be written as single character Unicode ۂ (U+06C2) or as a combination of two characters ہ and ٔ (U+06C1 and U+0654). In the same way, ۇ can be written in two different ways, i.e. using two characters' Unicode و with ٔ (U+0654 and U+0648) or with single character which is ۇ (U+0624). The extracted word list treats such variations as separate word based on the different Unicode values. Hence such issues need to be resolved using normalization of Urdu text. Few examples of normalization issues are shown in Table 2**.**

Table 2: Examples of normalization issues found and its replacement made

Issue	Replacement
جذبۂ (U+062C U+0630 U+0628 U+06C1 U+0654)	جذبۂ (U+062C U+0630 U+0628 U+06C2)
جرأت (U+062C U+0631 U+0627 U+0654 U+062A)	جرأت (U+062C U+0631 U+0623 U+062A)
لکم (U+0644 U+0643 U+0645)	لکم (U+0644 U+06A9 U+0645)
مشکوۃ (U+0645 U+0634 U+06A9 U+0648 U+0629)	مشکوٰۃ (U+0645 U+0634 U+06A9 U+0648 U+0670 U+06C3)

3.3.2. Aerabs. The extracted word list also contains words which are treated as separate words because of aerab attachment. In Urdu, aerabs are optionally used to give pronunciation guidance of same word but written/used in different contexts. Urdu has a variety of words which are written using same character sequence, but have different meaning based on the context in which they appeared. Such words have different phonetic behavior which is indicated using aerab. In Urdu writing styles, usually aerabs are not used and such words are separated using the context in which they are appeared, e.g. جگ can be used in two different contexts i.e.جُگ (\jʊg\) and جگ (\jæg\), but such words are normally written without aerab i.e. جگ. Based on this analysis, aerabs are removed from corpus.

3.3.3. Space Insertion and Omission. Space insertion and space deletion issues also exist in this corpus which are handled separately. Normally, in Urdu, space is not properly inserted between the words. The space deletion issues deal with the cases where space will be deleted inside compound words and if required Zero Width Non-Joiner (ZWNJ) will be used to preserve the valid word shape. In Urdu, words such as کم فہم, the ligature which ends with joiner will be attached with next ligature if space is removed and ZWNJ is not inserted between ligatures. The typists normally add space between ligatures which caused segmentation of a compound word into two words. This is due to the unfamiliarity of ZWNJ(U+200C) and unavailability of this character on keyboard. Therefore, to address this issue, a separate list is prepared which contains the compound words having space between ligatures. This space is automatically replaced with ZWNJ during cleaning process.

The following categories of Urdu compound words are also identified which are resolved automatically. The extracted word list is analyzed and different category of compound word indicators are extracted and against each category, the solution is implemented.

The compound words which are joined with ال are handled separately. All the words are extracted from the corpus, which start with ال prefix. This extracted list is manually analyzed and words are finalized, which will be joined with previous word in the corpus e.g. السلا ،الحق ،النفس. There are some words in Urdu which start with ال but these are independently valid words, e.g. الٹ ، الگ, hence all such cases are removed from the extracted list. The finalized ال word list is resolved in such a way that any word exists in this list is attached with previous word by deleting the space between them and if the previous word ends with joiner then ZWNJ is inserted e.g. بین الاقوامی is replaced with بین‌الاقوامی.

The compound words which contain زیر اضافت (Zeer-e-Azafat) i.e. ِ between sub-words are also handled automatically. The word which contains زیر اضافت (Zeer-e-Azafat) at end is automatically attached with next word in the corpus e.g. ادبِ لطیف. The attachment is done in such a way that if the word containing زیر اضافت (Zeer-e-Azafat) ends with joiner then ZWNJ is added between indicated word and next word.

Table 3: Examples of seen Zeer-e-Azafat in corpus

Issue	Replacement
ادبِ لطیف	ادبِ لطیف
فنونِ لطیفہ	فنونِ لطیفہ
جنگِ آزادی	جنگِ آزادی
حکومتِ ہند	حکومتِ ہند

The word which contains یائےاضافت (Yay-e-Azafat) at end can be sub-word of compound word, e.g. دریائے is sub word of دریائےراوی . Therefore a complete word list is extracted from the corpus, which contains یائےاضافت (Yay-e-Azafat) at end. This list is manually analyzed and finalized. There are Urdu words which

end with یائےاضافت (Yay-e-Azafat) but these are not part of compound word e.g. چھتیائے. Therefore, all such words are removed from the extracted list. This یائےاضافت (Yay-e-Azafat) word list contains all those sub-words which will be joined with the next word in the corpus by deleting the space. Some examples of یائےاضافت (Yay-e-Azafat) words are given in Table 4.

Table 4: Examples of Yay-e-Azafat found in the corpus

Issue	Replacement
دریائے راوی	دریائےراوی
دنیائے طب	دنیائےطب
اشیائے تجارت	اشیائےتجارت
دریائے کاویری	دریائےکاویری

Space insertion issues also exist in the corpus which are discussed below. The Urdu words which end with ء (HAMZA) must be separated with a space. As ء (HAMZA) is a non-joiner therefore usually space is not inserted after ء (HAMZA) to type next word. The ء (HAMZA) is a clear indicator of word boundary, e.g. شرکاء and اطباء therefore space is inserted after ء (HAMZA) to separate the next word.

The special symbols and punctuation marks are also handled automatically in such a way that space is inserted before and after special and punctuation symbols so that these cannot be attached with any Urdu word.

Normally space is not added between Urdu word and digit (Latin or Urdu digit) e.g. گھنٹے8. To resolve this issue, the corpus is processed and space is inserted between digits and Urdu character/letter. In the same way, the Latin word and Urdu words are not separated using space e.g. کا txt. Therefore the space is inserted between Latin and Urdu letters. Some example of such space insertion issues are given in Table 5.

Table 5: Examples of space omission in corpus as in case of numerals and Urdu text

Issue	Replacement
کےUKOU	کے UKOU
۷۵فیصد	۷۵ فیصد
اگست۱۹۴۷ء	اگست ۱۹۴۷ء
سے200مائیکرو	سے 200 مائیکرو

After careful analysis, it has been observed that the suggested solution to address these cleaning issues must be applied in an order so that proper words can be extracted. Therefore the order of solutions which are applied in automatic cleaning application is listed in Table 6. The automatic cleaning application is developed in such a way that each step is performed separately on complete corpus and then the next step is performed.

Table 6: Automatic cleaning process

Sequence No	Step
1	Remove space between "ال" words list and previous word in the corpus and add ZWNJ where required.
2	Apply Normalization
3	Automatically insert ZWNJ between the words by using the cleaned list of ZWNJ insertion between words
4	Add space between special symbols and punctuation marks but not within Latin words
5	Join word with next word which exists in یائےاضافت (Yay-e-Azafat) word list
6	Join word with next word which exists in زیر اضافت (Zeer-e-Azafat) word list
7	Remove All Aerab
8	Separate Latin digits from Urdu
9	Separate Urdu Unicode from Latin characters
10	Separate Urdu digits(۰-۹)from Urdu characters using space
11	Add space after ء (HAMZA)

3.4. Poetry Corpus Cleaning

Same as done for prose corpus, the poetry corpus is processed to remove HTML tags, URLs, and Non-Unicode letters. The manual analysis of the corpus shows that the poetry corpus cannot be cleaned using automatic cleaning application. Therefore, poetry books are cleaned manually using following cleaning guideline.

1. Introductory section is removed from poetry books to ensure only poetry text in the book.
2. The prose portion having dedication of the respective poem to someone is also removed.
3. Extra symbols such as ***** are removed.
4. Carriage return is inserted after each Verse (مصرع) and Couplet (شعر) so that these can be separated automatically.
5. Footnote is also removed.

After this manual cleaning, the poetry books contain only poetry text so that these can be further processed to extract the poetry N-grams.

3.5. N-grams Extraction

The N-grams give useful information of corpus which can be used in different NLP application. In this paper, the N-grams are extracted from prose and poetry corpora separately. N-grams are extracted at unigram, bigram and trigram levels for words and ligatures.

4. Results

The crawled corpus is categorized into poetry and prose genres. After manual corpus cleaning, the poetry corpus information such as number of books, poets, verses, words and unique words is given in Table 7. A total of 309,486 words are collected from poetry.

Table 7: Poetry corpus size

Number of Books	507
Number of Poets	331
Number of Verse	304,124
Total Words	309,486
Unique Words	42,883

The prose is manually classified into 18 sub-domains including articles, biography, character representation, culture, foreign literature, health, history, interviews, letters, magazines, novels, plays, religion, reviews, science, short stories, travel and Urdu literature The corpus is automatically cleaned using process discuss above. The number of books, words and unique words of each domain are given Table 8. A total of 37,680,293 words are collected from prose domain.

The N-grams are extracted from cleaned poetry corpus. The number of each computed N-grams are given in Table 9.

The automatically cleaned prose corpus is further processed to compute N-grams. Two different types of N-grams are computed from prose; (1) N-grams computed from prose and (2) N-grams computed from each classified sub domain of prose. The ligature N-grams and word N-grams are computed for each category of N-grams. The information about word N-grams and ligature N-grams computed from complete prose are given in Table 10 and Table 11 respectively. The domain wise information about word n-grams is given in Table 12.

Table 8: Domain wise corpus distribution of prose

Domain	Books	Total Words	Unique Words
Articles	59	1,645,456	45,023
Biography	34	872,350	30,142
Character Representation	26	643,661	23,923
Culture	5	245,557	13,921
Foreign Literature	27	485,311	17,895
Health	6	116,952	9,861
History	10	675,753	30,986
Interviews	12	6,84,776	20,955
Letters	7	2,91,666	18,221
Magazines	42	1,975,053	70,292
Novels	50	2,175,922	33,354
Plays	16	545,565	16,486
Religion	255	19,105,682	135,329
Reviews	51	2,058,784	48,667
Science	11	309,559	19,840
Short Stories	176	4,135,362	52,120
Travel	21	995,746	30,804
Urdu Literature	53	1,693,580	43,313

Table 9: Poetry corpus N-grams

Unigram	Bigram	Trigram
42,883	659,988	1,567,956

Table 10: Words N-grams

N-grams	Cleaned	Uncleaned
Unigram	239,580	498,916
Bigram	3,879,470	4,654,178
Trigram	13,109,156	14,649,881

Table 11: Ligatures N-grams

N-grams	Count
Unigram	91,665
Bigram	1,453,253
Trigram	7,038,582

Table 12: Domain wise N-grams

Domain	Unigram	Bigram	Trigram
Articles	45,023	455,318	1,004,308
Biography	30,142	274,156	560,093
Character Representation	23,923	210,499	419,477
Culture	13,921	96,900	171,466
Foreign Literature	17,895	151,985	299,163

Health	9,861	43,488	69,453
History	30,986	243,544	471,659
Interviews	20,955	203,657	446,315
Letters	18,221	122,445	211,417
Magazines	70,292	897,027	2,230,602
Novels	33,354	462,767	1,189,246
Plays	16,486	139,518	269,202
Religion	135,329	1,779,388	6,179,124
Reviews	48,667	522,772	1,175,068
Science	19,840	126,793	221,147
Short Stories	52,120	797,386	2,117,611
Travel	30,804	308,919	645,079
Urdu Literature	43,313	459,695	1,012,421

5. Discussion and Future Work

To generate the representative N-grams of Urdu corpus, 37 million words Urdu corpus is collected and categorized into two main domains. The automatic cleaning of prose resolves much of the word segmentation errors in less time. The extracted N-grams from prose give reasonable words and ligatures contextual information, but still there are some low frequent errors which require a manual cleaning pass. The higher order N-grams such as 4-grams and 5-grams are also useful for advance NLP applications such as machine translation system. These reported as future work.

The word level cleaning of the poetry is not performed. Hence, to have more accurate N-grams, poetry corpus needs to be cleaned so that proper word boundary can be defined. In future, the analysis of the poetry corpus will be carried out and some semi-automatic way of cleaning poetry corpus will be defined.

For future work, collected text corpus would be extended to annotate it with POS tags so that POS tagged N-grams will be extracted and reported.

6. Conclusion

In this paper, 37 million words corpus is processed and cleaned, a semi-automatic cleaning process is devised for such a large corpus. The categorization of the corpus into prose and poetry is also discussed. To ensure the diversity of the text corpus and to extract domain-specific N-grams, the prose is further categorized into 18 different domains. In future, the POS tagged layer will be added to generate the POS tagged N-grams. These N-grams can be used in any Urdu Natural Language Processing and Information Retrieval application. The presented Urdu books N-grams are publically available at: http://cle.org.pk/clestore/cleurdungrams.htm

7. Acknowledgement

This work has been conducted through the project, Urdu Nastalique OCR supported through a research grant from ICTRnD Fund, Pakistan.

8. References

[1] S Urooj, S Hussain, F Adeeba, F. Jabeen, R. Parveen, "CLE Urdu Digest Corpus", in Proc. *Conference on Language and Technology* (CLT12), Lahore, Pakistan. 2012

[2] M Ijaz, S Hussain. "Corpus Based Urdu Lexicon Development", in Proc. of *Conference on Language Technology (CLT07)*, University of Peshawar, Pakistan, 2007.

[3] G. Kennedy. (1998). "An introduction to corpus linguistics". *Addison Wesley Longman* Ltd. 1998

[4] E Darrudi, MR Hejazi, F Oroumchian. "Assessment of a modern farsi corpus." in *proc. of the 2nd Workshop on Information Technology & its Disciplines (WITID)*. 2004.

[5] Ale. Abolfazl, A. Hadi, D. Ehsan, R. Masoud, O. Farhad, "Hamshahri: A standard Persian text collection." *Knowledge-Based Systems* 22.5 (2009): 382-387.

[6] Y Lin, JB Michel, EL Aiden, J Orwant, "Syntactic annotations for the google books ngram corpus." In *Proc. ACL 2012 System Demonstrations*. Association for Computational Linguistics, 2012.

[7] T. Eckart, F. Alshargi, U. Quasthoff, D. Goldhahn. "Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin." In *Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools Workshop Program*. 2014.

[8] F. Mayer, Corpus Linguistics: Introduction. (1st edition), *Cambridge University Press*, 2002, Retrieved (06, 25, 2012).

[9] Urdu Library, Available: http://kitaben.urdulibrary.org/AllBooks.html