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, aimie.akram@kics.edu.pk

Availability

The presented Urdu books N-grams are publically available at:
http://cle.org.pk/clestore/cleurdungrams.htm

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 are reported. These corpora are extracted from Urdu magazine and news.

N-gram 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.

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. 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. 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.

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.

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.

Analysis showed that the extracted N-grams on uncleaned/raw 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 before extraction of N-grams. To aid the cleaning process for Urdu text, a cleaning tool 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.

Normalization: The Urdu words which can be written using different sequence of Unicode characters also exist in the corpus e.g. ؤ can be written as single character Unicode ؤ (U+06C2) or as a combination of two characters و and ٔ(U+06C1 and U+0654)

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 جَگ(\jæg\), but such words are normally written without aerab i.e. جگ Based on this analysis, aerabs are removed from corpus.

Space Omission: In Urdu, words such as کم فہم , the ligature which ends with joiner will be attached with next ligature if space is removed, therefore to maintain the shape a space is inserted in it making it two words instead of one. Instead of space it should be having ZWNJ inserted between ligatures. Similarly in other scenarios like Zeer-e-Azafat e.g ادبِ لطیف , Yay-e-Azafat e.g. دریائے راوی spaces are inserted unnecessarily which are replaced with ZWNJ.

Space Insertion: Sometimes due to non-joiner nature of characters a space is omitted while writing, all such issues are resolved by adding spaces at appropriate positions. For eg. As ء (HAMZA) is a non-joiner therefore usually space is not inserted after ء to type next word. Similarly space is inserted before and after special and punctuation symbols so that these cannot be attached with Urdu words. In same way space is not added between Urdu word and digit (Latin or Urdu digit) e.g. 8گھنٹے . Or between Latin words and Urdu words e.g.کا txt. Therefore the space is inserted in such cases.

N-gram Extraction

The N-grams give useful information of corpus which can be used in different NLP application. Here, 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.

Results

The crawled corpus is categorized into poetry and prose genres.

Poetry:

After manual cleaning of poetry corpus , the corpus information such as number of books, poets, verses, words and unique words is given in Table 1. The N-grams are extracted from cleaned poetry corpus. The number of each computed N-grams are given in Table 2

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

Table 1: Poetry Statistics

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

Table 2: Poetry Corpus N-grams

N-grams	Cleaned	Uncleaned
Unigram	247,409	498,916
Bigram	4,117,209	4,654,178
Trigram	14,692,464	14,649,881

Table4: Words N-grams Count

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

Table5: Ligature N-gram Count

Prose:

The prose is manually classified into 18 sub-domains including articles, biography, character representation, culture, foreign literature, health and Urdu literature etc. The corpus is automatically cleaned. The number of books, words and unique words of each domain are given Table 3.

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 4 and Table 5 respectively. The domain wise information about word n-grams is given in Table 6.

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 3: Domainwise Corpus Distribution of Prose

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

Table 6: Domain wise N-grams

Conclusion

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. process is devised for such a large corpus.