Acta Scientific COMPUTER SCIENCES

Volume 3 Issue 10 October 2021

Research Article

Pattern Defined Corpus Construction: A Complete Training to Building the Corpus

Sayed Majid Ali Shah*, Zeeshan Bhatti, Zulfiqar Ali Bhutto and Kamran Taj Pathan

Dr. A.H.S. Bukhari Institute of Information Communication and Technology, University of Sindh, Jamshoro, Pakistan

***Corresponding Author:** Sayed Majid Ali Shah, Dr. A.H.S. Bukhari Institute of Information Communication and Technology, University of Sindh, Jamshoro, Pakistan.

Received: August 22, 2021

Published: September 14, 2021

Abstract

Corpus is considering the mandatory component required for the processing of any language to building the Natural Language Processing (NLP) applications that perform the tasks, particularly the language analysis, manipulation, and information retrieval. In this article, a procedure has been discussed and illustrated for the constructions of the corpus. This study illustrates the training about the constructions of the corpus in any language. Numerous approaches have been sketched through different perspectives of the language with the support of the Sindhi language. The applications including machine translation, spell checking, grammar checking, parts of speech tagging, named entity recognition and word identification also have been addressed. The text has been taken from various digital sources such as newspaper websites, blogs, e-books, and magazines. The procedural models also have been demonstrated for the NLP applications by using the corpus.

Keywords: Natural Language Processing (NLP); Corpus; Sindhi Language

Introduction

A large collection of text in a structured form known as Corpus. It is the branch of linguistics. A scientific study of language considers linguistics such as analysis of the language, change of language behavior [1,2]. The corpus allows the researchers' individuals to build the new algorithms by the using corpus. For computational linguistics, it is mandatory to obtain the script in the digital form as E-text. The text has been taken by multiple sources from the internet [3]. The Sindhi language is the Indo Aryan Language, very rich in a morphological structure having fifty-two alphabetical characters. Sindhi is the official language of Sindh, Pakistan as well as in India. The written script of Sindhi starts from the Right-to-Left direction similar to the Arabic language. About 59 million Sindhi native speakers are found across the world [4,5]. There is no such a Sindhi corpus publicly available (to the best of my approach). However, with the huge number of Sindhi native speakers, it is mandatory to construct the corpus that represents the Sindhi language for the computational linguistics process of the language.

Advantages: The corpus facilitates the individuals, researchers who work in the field of Sindhi NLP advanced applications development by using artificial intelligence such as text to voice, voice to text recognition. It also helps Sindhi spell checking, Sindhi Optical Character Recognition (OCR).

Related work

Myanmar (Burma) National Language corpus was constructed based on the XML tagging. There were 109 generic languages collected. The purpose was to develop a custom tag-set corpus to facilitate generic languages in NLP applications. It was manually tested. The text was collected from various newspapers and e-books [6]. Another hand, the Hindi corpus was built at IIT Bombay with the aims of machine translation. There was a parallel corpus, which covers the 174k sentences in Hindi to English. The auto-linguistics tools were applied that are available on the internet with filter techniques [7]. Furthermore, an Italian corpus was developed based on parallel corpus for the machine translation from Italian to English.

The SQL was used in this translation corpus. It contains 4.6 million texts in the form of a bilingual corpus [8]. The British National Corpus (BNC) was built in the English language which covers 100 million texts. The SQL database was used. This study aimed to facilitate the English language in computational linguistics and also for advanced NLP application development [9]. Moreover, the Sindhi Language Corpus (SLC) was construct based on XML custom tagging with MySQL database using a rule-based approach. The XML documents were created for the facilitation of the machine learning process and also for the researchers to use this corpus and built rules based on the application requirement [10]. Various corpora have been discussed such as MNC, BNC, ANC they used distinct techniques, but the proposed research based on SLC differs from the existing corpora. The proposed corpus covers the Sindhi language, it is different from the previous corpus. other corpora used the XML, or SQL, or auto filter online linguistics tools, but the SLC used the MySQL database for the storage with apache server and for the linguistics analysis, the built-in GUI have been used with SQL queries. The processed text has been forwarded by creating the GUI-based corpus software using NetBeans.

Procedural models

Initially, must select the language in which want to construct the corpus and planned the source of language generation in the form of a digital script of that language. The text should be processed for further actions which include the operations as paragraph or text extraction into sentences as sentence segmentation. After that the text breakdown into tokens as tokenization. However, all the text should be stored in a database.

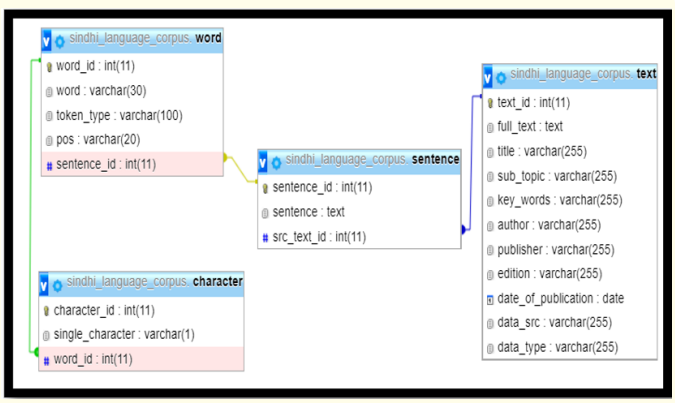

Figure 1: Relationships between corpus tables.

In the database section, the tables named include complete text with the attributes such as text_id, full_text, text_source, source_type. This is the main table of the corpus. The attribute text_id holds the auto number of the rows as records, the full_text attribute contains the actual text of the paragraph/lesson or chapter. The text_source attribute shows where comes the data? It holds the URL link, where full_text has been taken. The source_type covers the information of the source either it is a newspaper, or book, or blog, or magazine, or any other website. The other table is named as sentences. In the sentence table the attributes include sentence_id, full_sentence, full_text_id. The sentence_id attribute is auto-generation based on entry records. The full_sentence attribute holds the complete sentence which was extracted from full_text while the full_text id attribute is the foreign key that exists in the sentencing table. The table named tokens based on the attributes, token_id, token, token_type, sentence_id. The attribute token_id considers the auto-generate number key and is used for the primary key within a table. The token column holds the single word extracted from the sentencing table. The token_type column covers the information of a token or holds the information of word identification, it is also useful for the named entity recognition. The token_type holds the suitable information relevant to a particular token either it is a personal name or city name or a common word. While the sentence_id is used as a foreign key of the tokens table. The character table is also used to store the characters and those extracted from the tokens (words). however, the other attributes in the text table are used for the metadata for avoiding copyright purposes such as date of publication, publisher, edition, etc. Four tables are having a relationship to each other and are illustrated in figure 1. The words used for the named entity recognition and words identification purpose, parts of speech tagging, frequencies of the words calculate for the next word prediction with the algorithm. The sentences are used for the word tokenization purpose and n-gram technique. N-gram frequencies calculation for the sentence prediction and gives more accuracy.

Figure 2 shows the text processing model for corpus construction, start with the input text, then it will segment into sentences. The words will extract from the sentences as it has been already discussed. The sentences are stored in the sentences table in a database, while the words will not directly be stored in the database because there is a Sindhi lexicon predefined where each word matched to the Sindhi lexicon for the identification process and assign the label to each word (token) as token_type and Sindhi Parts

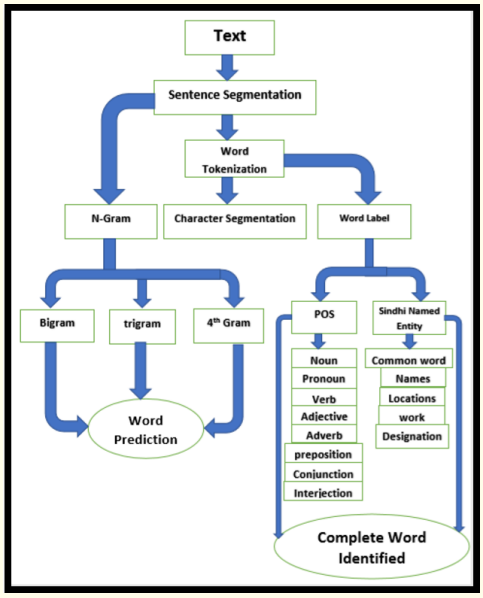

Figure 2: Corpus Text Processing Model.

of Speech (POS). The Sindhi POS covers the grammatical part of the token, though, the token_type covers the words identification task and also the named entity recognition. The N-gram is extracted from the sentences.

Results and Discussion

Finally, the corpus has been constructed based on the Sindhi language. And illustrated in figure 3.

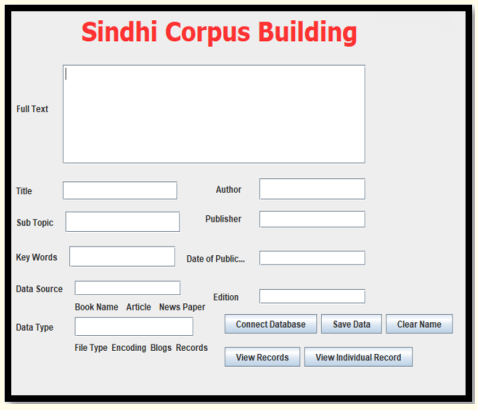

Figure 3: Graphical Interface of the corpus.

The full-text area intakes the input from the user. Then enter the title of the text, the sub-topic is also added. And the keywords are mandatory to represent the text. The author name, publisher name, date of the publication, edition, data source, and data type are also the required fields and they also considered the metadata of the document while they used to avoid the copyright issue. After filling all mandatory text fields connect to the database and click on a save data button. The data will be saved into the database. furthermore, the stored data is illustrated in figure 4.

Figure 4: Text stored into a database.

The full text has been stored in the database along with the metadata of the document and is mentioned in figure 4.

Figure 5: Sentence Stored into the Corpus Database.

Figure 5 describes the sentences that have been stored into the corpus database and it has been extracted from the full text, while in the sentencing table there is src_text_id mentioned which shows the relationship to a particular text file.

word_id	word	token_type	pos	sentence_id
1150058	جي	عام لفظ	حرف جر	69851
1150057	حيدرآباد	شهر	اسم معرفہ يا خاص	69851
1150056	طرفان	عام لفظ	ظرف مڪان	69851
1150055	يونيورسٽي	علم پرائڻ جي جڳهه	اسم تصغيره يا عام	69851
1150054	خان	دالو	اسم تصغيره يا عام	69851
1150053	آءٿا	unknown	unknown	69851
1150052	اڳ	عام لفظ	حرف جر	69851
1150051	کان	عام لفظ	حرف جر	69851
1150050	ان	عام لفظ	ضمير اشارو	69851
1150049	ويو	حڪم	فعل	69851

Figure 6: Single Word Stored into the Corpus Database.

Figure 6 demonstrates the single word as a token stored in the database. There is token_type attribute which shows the type of the token in simple terms, to shows the identification of a word in Sindhi language either it is name or city or work or common word, etc. Another hand, there is a pos attribute that has stored the information of Sindhi grammatical while the sentence_id shows the relationship that the word has been extracted from a particular sentence.

Conclusion

Corpus considers the core component to build the NLP applications. A process has been defined for corpus construction. Only the Sindhi language is addressed in this study. The database considers the main part of the corpus. The database development with corpus scenario is illustrated in discussed in detail. The text processing model describes which illustrates how the text is processed from a paragraph to a single token and a single character. after the text processing, the data is inserted into the database by the graphical interface for the Sindhi corpus. The inserted text is also shown in various figures as text, sentences, single token. However, the token_type used for the Sindhi named entity recognition purpose and Sindhi pos fetched by using Sindhi lexicon. The 1.2 million Sindhi texts were collected from internet sources.

Bibliography

1. Meyer C F. "English corpus linguistics: an introduction". Cambridge: Cambridge University Press (2002).
2. Powell C and R Simpson. "Collaboration between corpus linguistics and digital librarians for the MICASE web search interface". In R. Simpson and J. Swales (2001): 32-47.
3. Ismaili I A., *et al*. "Design and Development of the Graphical User Interface for Sindhi Language". arXiv preprint arXiv (2014): 1401.1486.
4. Bhatti Z., *et al*. "Phonetic based soundex and shapeex algorithm for sindhi spell checker system". arXiv preprint arXiv (2014): 1405.3033.
5. Bhatti Z., *et al*. "Word segmentation model for Sindhi text". *American Journal of Computing Research Repository* 2.1 (2014): 1-7.
6. Ko W K and Phyo T Z. "Selection of XML tag set for Myanmar National Corpus". In Proceedings of the 6th Workshop on Asian Language Resources (2008).
7. Bojar O., *et al*. "HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation". In LREC (2014): 3550-3555.
8. Zanettin F. "CEXI: designing an english Italian translational corpus". In Teaching and Learning by Doing Corpus Analysis. Brill Rodopi (2002): 327-343.
9. Meyer C F. "English corpus linguistics: An introduction". Cambridge University Press (2002).
10. Bhatti Z and Shah M. "Sindhi Text Corpus using XML and Custom Tags". *Sukkur IBA Journal of Computing and Mathematical Sciences* 2.2 (2018): 30-37.

Volume 3 Issue 10 october 2021