JOURNAL oF CRITICAL REVIEWS
ISSN- 2394-5125 VOL 7, ISSUE 09, 2020

ON CREATION OF DOGRI LANGUAGE CORPUS

[1]Sonam Gandotra, [2]Bhavna Arora
[1]Ph.D, Department of Computer Science & IT, Central University of Jammu.
[2]Assistant Professor, Department of Computer Science & IT, Central University of Jammu

AbstractThe pre-requisite for any Natural Language Processing (NLP) task, is the corpus. Corpus is defined as a large collection of structured text. Dogri is one of the official languages of India but is under-resourced in terms of computational resources needed for any NLP task. This paper proposes a methodology to construct a standard corpus which can be used for performing various language processing tasks like stemming, part-of-speech tagging, information retrieval, etc. The digitized text required for creating the corpus is not available due to the scarcity of online resources containing Dogri text. The only online source which is available is the Dogri Newspaper "Jammu Prabhat". Hence, the text is to be extracted from portable document formats (pdf) of that newspaper which are first converted to images before extraction of the text. To achieve this, an open-source tool-Tesseract is used for extracting the text from images. The methodology that is used for the corpus creation of Dogri Language is discussed in detail in the paper. The challenges faced during the research and the acquired results have also been discussed.

Index Terms— Corpus, Tesseract-OCR, Dogri Language, Language Resource.

I. INTRODUCTION

L ANGUAGEis one of the fundamental traits of human behavior which enable us to express and communicate with people across the world. It can either be in spoken or written form [1]. With the advancement in technology, the researchers are trying to build tools which can understand the natural language and make technology more accessible to the people. Natural Language Processing (NLP) has been an active area of research with the objective of making computer able to process, understand, summarize, translate and draw inferences from the human natural text and language. Various linguistic resources are required for carrying out diverse NLP tasks like machine translation, automatic text summarization, sentiment analysis, named entity recognition, parsing etc. These linguistic resources mainly deal with processing of text in written form i.e. performing processes like tokenization, stemming, parsing etc. Due to progression in research wide number of languages are being digitalized and efficient tools has been developed for them[2]–[5]. But as far as Indian languages are concerned, it is still in its progressive phase due to the variability of the Indian languages and their complex structure. The linguistic resources required for the development of tools required for processing in NLP are also limited. Researchers are constantly working on providing the basic tools required for the processing of text in digital form so that better results could be achieved [6]–[8].

In India, there are 22 official languages as defined in the eighth schedule of the Indian Constitution [9] and Dogri language being one of them. Dogri language belongs to the class of Indo-Aryan languages, which is spoken by about 5 million people in India and Pakistan but mostly has its dominance in the state of Jammu & Kashmir and parts of Himachal Pradesh and northern Punjab [10]. Dogri was originally written using the Dogri script [11] which has its resemblance with the Takri script, the script of royals of Himachal Region but now the language is written using Devanagari script. Dogri language is still in its developing form, as it has not been widely explored in area of computational linguistic and natural language processing. Thus, resulting in scarcity of online resources required for the processing of text.

Dogri is the native language of the people of J&K and in particular of the people of Jammu region. The digitization of this language for the creation of corpus is taken up as the researcher itself hails from this region. In this era of digitalization, it is important to take up every minor step which will bring the smaller of the part in race with the global forum. The creation of Dogri corpus will lead more researchers being able to develop the NLP tools for Dogri language which will further enhance the research for the language.

Corpus is defined as the large collection of data which can either be in written or spoken form. It is machine-readable and is used for the purpose of linguistic research. A standard corpus is required for the development of tools needed for the processing of language. There are various benchmark corpuses which researchers use for executing

NLP tasks in different domains like Brown Corpus[12], DUC (Document Understanding Conferences) [13], TAC (Text Analysis Conference)[14], CNN dataset, Reuters dataset etc. These corpuses contain tokens, human generated abstractive as well as extractive summaries required for evaluation of the proposed summarization technique. These can be used for both single and multi-document summarization. These datasets are mainly developed for English, Chinese and other European languages. As far as Indian languages are concerned, Indian languages are less explored as compared to the foreign languages due to the lack of availability of resources required for processing the text. This scenario is changing rapidly through the efforts of the researcher's community, as various linguistic resources are being developed which are at par with the resources developed for foreign languages. Few examples of resources for text are Parallel Corpora, Ontology & Word-Net, Online Vishavkosh, Text Corpora etc. [15], [16] and for speech are Speech Corpora, Semi-automatic annotation tool for Speech Corpora [17].

This paper aims at developing a corpus for Dogri language. The created corpus contains news items covering the domain of sports, education, politics, entertainment from the only Dogri newspaper "Jammu Prabhat" [18]. The idea behind choosing the newspaper is to cover maximum words from all the domains. As the data is not available in digital form and the only available resource are the images of the text. Thus, in this work, an open source tool i.e. Tesseract [19] is used for extracting the text and making a corpus from the available resources. The paper begins with the introduction of the linguistic resources required for performing NLP tasks. A brief introduction of Dogri language along with the status of on-going research in the field of NLP for Dogri language is presented in Section I. The background work that has been carried out previously for under-resourced languages and the different techniques used by researchers to extract text from images are also discussed in Section II. In Section III, the proposed methodology for creation of Dogri corpus and the need for its development are also described. Section IV gives a detailed explanation of the methodology for corpus creation. The challenges which were faced during creation of the corpus are presented in Section V. Then, the paper discusses the statistics which were obtained during extraction of text from the newspaper and book pdf in Section VI. Finally, the paper concludes in Section VII along with the open challenges for further research in this area are also stated.

II. BACKGROUND

The task of identification of corpus or to get the corpus in digitised form is not a challenge anymore for the common languages having widespread usage and implementation. As large numbers of resources are freely available and now the researchers are concentrating on making improvements in the techniques which can be applied further to enhance the quality of research. Under-resourced languages are deprived from these basic resources like: corpus, annotators, part of speech tagger, stemmer, named-entity lists etc. These resources provide a base for developing efficient tools and applications which boost NLP research. Over the years the researchers have developed tools and techniques for extracting text from the images. The researchers are also working for developing the basic resources for the under-resourced languages. A few of these tools and the basic resources are discussed below as:

In an attempt of using Tesseract-OCR, Kumar [20] have used Tesseract-OCR for recognizing Gujarati characters. An already available trained Gujarati script was used for training the OCR. The results have been checked by using different font-size, font style etc. and the mean confidence come out to be 86%. Various neural network techniques have been proposed by the researchers for identification of text from the images. Word level script has been identified with the help of an end-to-end Recurrent Neural Network (RNN) based architecture which is able to detect and recognize the script and text of 12 Indian languages by Mathew in [21]. Two RNN structure was used each containing 3 levels of 50 Long Short-Term Memory (LSTM) nodes; one RNN is used for script identification and the other is used for recognition. Smitha [22] has described the whole process of extraction of text from the document images. The process is executed with the combination of Tesseract-OCR and imagemagick and then the results are compared with that of OCRopus.OCRopus is also an open source OCR developed for English language. But the results of the Tesseract-OCR and imagemagick combination are more promising as compared to the OCRopus.

Sankaran in paper [23] has proposed a recognition system for Devanagari script using RNN known as Bidirectional Long Short-Term Memory (BLSTM). The approach used by the author has eliminated the need for word to character segmentation thus reducing the word error rate by 20% and the character error rate by 9% as compared to the best OCR available for the Devanagari script. In [24], Chakraborty has developed a system for the visually-impaired person in which the text from the

images are extracted and then converted to a 6-dot cell Braille format. The scanned document images are first converted into grayscale images and then passed to the Tesseract for extraction of text from these scanned images of the document. The text is then processed by using API JOrtho for any errors in the extracted text. The prepared data is then converted to the Braille format using a defined set of rules. The application is useful and considered as the first of its type. The relative study of two optical character recognition tools i.e. Tesseract and Transym has been done by Patel in [25]. The later one is the open-source tool while the other is a commercial tool. The experiments have been carried out for extraction of vehicle numbers from the vehicle number plates. Tesseract comes out to be more powerful as compared to Transym. The accuracy for coloured images is 61% while for grayscale images it is 70% in case of Tesseract OCR whereas it is only 47% in Transym.

A complete framework for integration of Bangla script recognition support in Tesseract was developed by Hasnat in [26]. The entire task is divided into two phases: training data generation and test data processing. 340 characters including 50 basic, 10 vowel modifiers and 270 compound characters are considered for initial training of the Tesseract. Dependent modifiers are trained separately and combined with the basic units. Further, binarization, noise elimination, skew detection and correction and character segmentation are carried out for the testing phase. Kumar in [27] have given an idea of developing Dogri WordNet in terms of Hindi WordNet due to the lexical and structural similarity of the languages. It is considered as the first attempt in developing such a lexical resource which will boost the research further in the area of NLP as various language processing tools such as stemmer, parser, spell checker etc. could be developed with the help of this WordNet. The English WordNet and the Hindi WordNet has enhanced the quality of research and the development of this WordNet also aims for the same.

Urdu corpus with the aim of enhancing the researcher's interest in the area of NLP was developed by Humayoun in [28]. Two versions of the same corpus were created. The first version contains the words separated by space while in second version, proper word boundaries are tagged manually. Further, normalization, part-of-speech tagging, lemmatization, stemming, morphological analysis is applied on the both versions of the corpus for better results. 50 articles were taken into consideration from various fields for the creation of a benchmark corpus. Gupta has developed an automatic extractive text summarization system for Punjabi text in [29]. The author first created the corpus from local newspaper and then applies pre-processing and processing phases to the text for generation of the summary. It was also the first attempt by the authors to create such system which further enhances the research carried out in India in the field of natural language processing.

III. PROPOSED WORK

The proposed work aims on creation of the Dogri corpus. Dogri is the native language of the state of Jammu and Kashmir. It is spoken by about 5 million people. But the digitalization of language is still a major concern. Due to lack of digitalization, the basic resources have to be built from scratch. The need for corpus creation, Tesseract OCR and the methodology that have been adopted in building the final corpus is presented in the sub-sections mentioned below.

A. Need for Dogri Corpus Creation

Creation of corpus is the foremost requirement for any NLP task. The input data chosen for corpus creation must be diverse in nature, so that the results could be evaluated efficiently. Dogri language being new to the area of NLP, no such linguistic resources is available for researchers for evaluating their research or to build new tools. It is the prime need for any natural language processing task to have a corpus. So, this paper proposes to build the corpus for the same. The scarcity of linguistic resources required for development of NLP tools has motivated this research.

B. Tools Used

Following tools have been used for creation of the corpus:

1) Pdf Files: There is scarcity of digitized data for Dogri Language. In order to create a corpus, a vast and diversified data is required. For the want of diversified data, newspaper articles from the Dogri newspaper "Jammu Prabhat" and a textbook of curriculum of Dogri language are chosen to be included in the corpus. The pdfs of these newspaper articles and textbook are used as the prime source of data for text extraction. These pdfs are converted to images first for further processing.

2) Imagemagick: It is open source software which is used for converting, editing and manipulating images. This tool is used for enhancing the quality of the image to be given as input to the Tesseract-OCR.

3) Tesseract-OCR: Tesseract [19], is an open source OCR engine outsourced by Google© having the ability to recognize more than 100 languages and one feature of this tool is that it can also be trained to recognize new languages [30]. This tool has been used for extracting the text from the available images in this paper. The foremost

reasons for choosing Tesseract is that it is freely available open source technology. Secondly, it provides a vast domain of languages that can be recognized using this tool and if the results are not satisfactory, then one can train its own language simply by adding box files to the training data. Thirdly, it is known to be the most accurate OCR till now giving up to 99% accuracy.

4) Python: It is an open source programming language and is used for analyzing the extracted text. The statistics of the corpus created are calculated by making use of this language.

IV. METHODOLOGY FOR CORPUS CREATION

Natural Language Processing tasks are highly dependent on the corpus for proper processing of the text. There are many sources available from which the corpus is readily available, but for under-resourced like Dogri all the work has to be done from scratch. The methodology that has been adopted for the creation of the corpus is presented in Fig. 1.

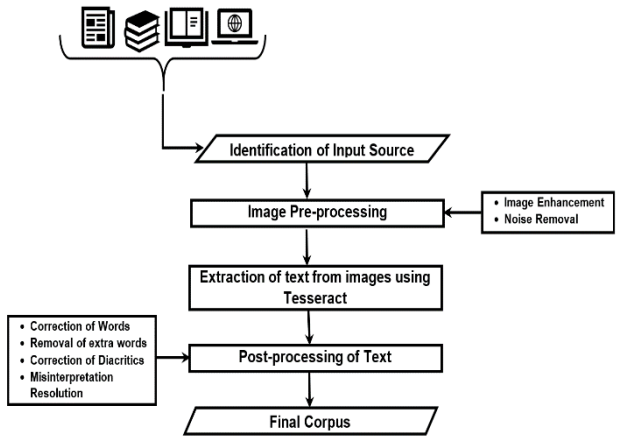


Fig. 1. Methodology Adopted for Corpus Creation

The detailed description of each of the above stated step is presented below:

A. Identification of the Input Source

From a variety of sources available e.g. novels, books, magazines, newspapers etc., newspaper images are chosen as the prime source for extracting text data. The main reason for choosing newspaper as an input source includes:

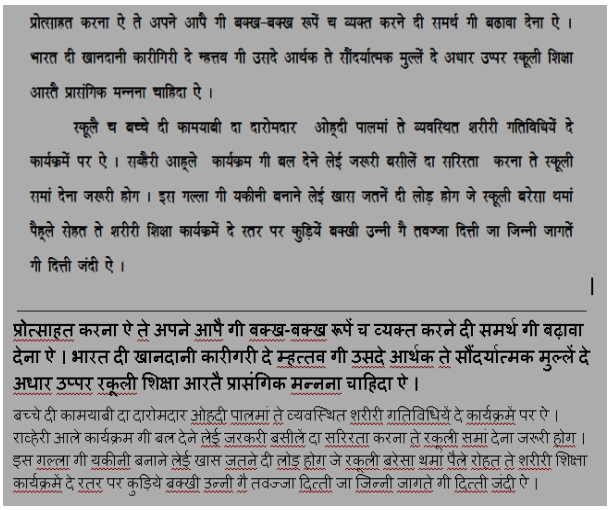
1) The variety of information present in them as they include items from all the spheres ranging from politics to entertainment to sports and education. Hence, providing a versatile corpus which can act as a standard for further NLP research.
2) The language used is understandable and identified by the regional community.
3) It is freely available and there is no need for further scanning as it is already in jpeg format.

B. Image Pre-processing

The available images are to be converted into png (portable network graphics) image format for better recognition of the text from images. As png formats allow lossless compression of the images in grey scale, thus enhancing the quality of images and making it suitable to be given as input to the Tesseract. The conversion of images from jpeg to png is done by using the imagemagick[31]. Imagemagick is an open source tool for converting, displaying andediting images. It is an inbuilt framework which deals with the binarization, noise removal, foreground detection and skew correction of the given image. Binarization, deals with the conversion of given coloured or grayscale document image into its bi-level representation [32]. The bi-level representation of the document image is then analysed for any unwanted noise in the image. After noise removal, the background and foreground of the image are separated for clarity in the given document image. If required, a skew is applied to the document image for further cleaning of the image [22]. All these sub-processes are part of the imagemagick framework which is taken care automatically. The converted and clearer image is then provided as input to theTesseract OCR.

C. Extraction of Text from Images using Tesseract-OCR

After successful conversion of the image from jpeg to png format, the image is fed as input to the Tesseract OCR for text recognition and extraction. Since, both Dogri and Hindi languages are written using the Devanagari Script, there is a structural similarity in the structure of basic alphabets and construction of words. Although being structurally similar, there is a huge difference in the vocabulary and context of the language which distinguishes both these languages. Tesseract software is trained on over 100 languages and for Dogri language; this paper proposes to use the Hindi training data because of the lexical similarity of the two. The Hindi training set is already available in Tesseract which contains the box file of the basic characters of the language and the word list of Hindi script. Using the Hindi training set on the corpus has yielded comparable results and is able to extract more than 80% of the text from the input images. The output of the PNG image given as input to the Tesseract is presented in the Fig.2 below:


प्रोत्साहत करना ऐ ते अपने आपै गी बक्ख-बक्ख रूपें च व्यक्त करने दी रामर्थ गी बढ़ावा देना ऐ । भारत दी खानदानी कारीगिरी दे म्हत्तव गी उसदे आर्थक ते सौंदर्यात्मक मुल्लें दे अधार उप्पर स्कूली शिक्षा आरतै प्रासंगिक मन्ननी चाहिदा ऐ ।

रकूलै च बच्चे दी कामयाबी दा दारोमदार ओह्दी पालमां ते व्यवस्थित शरीरी गतिविधियें दे कार्यक्रमें पर ऐ । राव्हेरी आह्ले कार्यक्रम गी बल देने लेई जरूरी बरीलें दा सरिरता करना ते रकूली रामां देना जरूरी होग । इस गल्ला गी यकीनी बनाने लेई खास जतनें दी लोड़ होग जे स्कूली बरेरा थमां पैह्ले रोहत ते शरीरी शिक्षा कार्यक्रमें दे रतर पर कुड़ियें बक्खी उन्नी गै तवज्जा दित्ती जा जिन्नी जागतें गी दित्ती जंदी ऐ ।

प्रोत्साहत करना ऐ ते अपने आपै गी बक्ख-बक्ख रूपें च व्यक्त करने दी समर्थ गी बढ़ावा देना ऐ । भारत दी खानदानी कारीगरी दे म्हत्तव गी उसदे आर्थक ते सौंदर्यात्मक मुल्लें दे अधार उप्पर रकूली शिक्षा आरतै प्रासंगिक मन्ननी चाहिदा ऐ ।

बच्चे दी कामयाबी दा दारोमदार ओह्दी पालमां ते व्यवस्थित शरीरी गतिविधियें दे कार्यक्रमें पर ऐ । राव्हेरी आले कार्यक्रम गी बल देने लेई जरूरी बसीलें दा सरिरता करना ते रकूली समां देना जरूरी होग । इस गल्ला गी यकीनी बनाने लेई खास जतनें दी लोड़ होग जे रकूली बरेसा थमां पैले रोहत ते शरीरी शिक्षा कार्यक्रमें दे रतर पर कुड़िये बक्खी उन्नी गै तवज्जा दित्ती जा जिन्नी जागते गी दित्ती जंदी ऐ ।

Fig.2. PNG image given to the Tesseract OCR and the text extracted after Tesseract Processing

D. Post-processing of Text and Creation of Corpus

The text extracted from the Tesseract OCR is not 100% accurate. So, it requires some manual processing for creation of the corpus. The manual processing includes separating words, identifying the end of line and correction of words which are not correctly identified by the Tesseract engine. An example of these pre-processing terms is presented below:

1) Removal of extra symbol: During the text extraction phase, the Tesseract OCR insert symbol such as "|" and "||", which does not demark the sentence rather add anomaly to it.
2) Misinterpretation of "स" with "र": After careful analysis of the extracted text, it is found that the Tesseract OCR is not able to completely recognize the character "स" of the Dogri text and misinterpret it as another Dogri language character "र".
3) Correction of diacritic "ै" with "े": Also, it is found that the Tesseract OCR, interpret the diacritic "ै" as "े" in few instances such as it recognizes the word "बारै" as "बारे", "लैन" as "लेन", "खबरै" as "खबरे" etc. And in certain instances, completely ignores the diacritic "ै".
4) Correction of words: The misinterpreted words were corrected manually to ensure the correctness of the corpus created.

After the manual processing, 200 documents have been identified to be included in the corpus. Different documents from different domains have been chosen whose text can be included in the final corpus. Thus, the corpus that is created from a vast variety of input text can be used for various NLP tasks like stemming, tokenization, translation, summary generation etc. The main areas from which document are selected includes the following: Sports, Politics, Entertainment, Tourism, Science & Technology, Religion, Miscellaneous etc.

V. CHALLENGES IN CREATION OF DOGRI CORPUS

Following are the challenges that are faced during the creation of Dogri corpus from images:

1) Limitation of resources: The major challenge for creation of corpus is that there is limited digital data available for Dogri language. The only digital data available is the Dogri language regional newspaper titled "Jammu Prabhat". The other non-digital resources of Dogri text includes text books, literature and magazines but is still very limited.
2) Multi format data: The content that is available cannot be copied directly into text form as the digital form contains images of the text.
3) Text conversion: Since the data that is available is not in a format that can be used to create corpus, the process involves additional overhead of text conversion. The text is scanned first and converted into either pdf or jpeg format. Direct extraction of text from the available resources is not possible. So, an OCR is required for extracting the text from the available sources.

VI. RESULTS AND DISCUSSIONS

The text is extracted from the images of the Dogri newspaper "Jammu Prabhat". The created corpus contains 200 documents with a total of 23,398 sentences with 4,72,271 tokens and 24,893 unique tokens. A Named entity list is also created which contains 593 tokens. Also, the corpus is UTF-8 encoded and is in text file format which can easily be used for further processing. The screenshot of the created corpus along with corpus statistics is presented in Fig. 3 below:

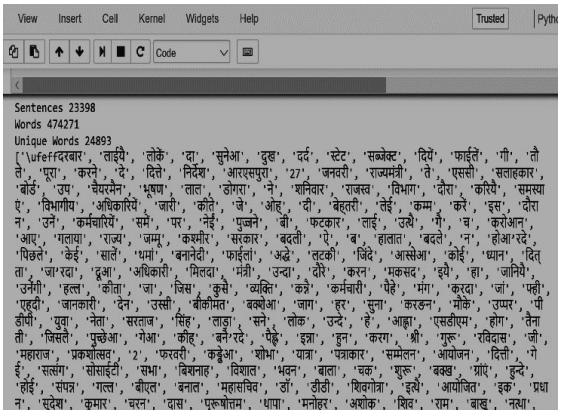


Fig. 3. Screenshot of the corpus statistics

VII. CONCLUSION

The work presented in this paper is part of the on-going research for building an automatic summary generator for Dogri text. This paper presents the methodology to build the Dogri language corpus that can be used for a variety of natural language processing tasks. The various tasks include stemming, tokenization, lemmatization, POS tagger, summary generation etc. For extraction of text, Tesseract which is an open source tool is used. The results obtained of this research can be further improved by using the exclusive Dogri language training.

REFERENCES

[1] J. Allen, Natural language understanding, 2nd ed. Pearson Education, 2002.

[2] R. Socher, J. Bauer, C. D. Manning, and A. Y. Ng, "Parsing with Compositional Vector Grammars," in ACL 2013, 2013.

[3] C. L. Philip Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," Inf. Sci. (Ny)., vol. 275, pp. 314–347, 2014.

[4] M. F. Porter, "An algorithm for suffix stripping," Program, vol. 14, no. 3, pp. 130–137, 1980.

[5] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network," in HLT-NAACL 2003, 2003, pp. 252–259.

[6] R. Puri, R. P. S. Bedi, and V. Goyal, "Automated Stopwords Identification in Punjabi Documents," An Int. J. Eng. Sci., vol. 8, no. June 2013, pp. 119–125, 2013.

[7] R. Puri, R. P. S. Bedi, and V. Goyal, "Punjabi stemmer using Punjabi wordnet database," Indian J. Sci. Technol., vol. 8, no. 27, 2015.

[8] A. Ramanathan and D. D. Rao, "A Lightweight Stemmer for Hindi," Proc. EACL 2003 Work. Comput. Linguist. South Asian Lang., pp. 43–48, 2003.

[9] "Abstract of Speaker's Strength of Languages and Mother Tongues-2011," 2011.

[10] S. R. Sharma, Encyclopaedia of Teaching Languages in India, v. 20. Anmol Publications, 1992.

[11] A. Pandey, L2/15-234R: Proposal to encode the Dogra script in Unicode. 2015.

[12] W. N. Francis and H. Kucera, "CoRD | The Brown Corpus (BROWN)," 1964. [Online]. Available: http://www.helsinki.fi/varieng/CoRD/corpora/BROWN/.

[13] NIST, "Document Understanding Conferences - Past Data," 2002. [Online]. Available: https://duc.nist.gov/data.html.

[14] NIST, "Text Analysis Conference (TAC) Data," 2008. [Online]. Available: https://tac.nist.gov/data/.

[15] S. Lata and S. Chandra, "Development of Linguistic Resources and Tools for providing multilingual Solutions in Indian Languages – A Report on National Initiative."

[16] P. CDAC GIST, "Indian Language Technology Proliferation and Deployment Centre - Home," Ministry of Electronics & Information Technology, MeitY, Govt. of India, 2016. [Online]. Available: http://tdil-dc.in/index.php?lang=en.

[17] L.-I. CIIL, "LDC-IL," Department of Higher Education, Ministry of Human Resource Development, Govt. of India, 2016. [Online]. Available: http://www.ldcil.org/resourcesSampleSpeechCorp.aspx.

[18] "Jammu Prabhat: First and Only Dogri News Paper, Dogri Newspaper." [Online]. Available: http://www.jammuprabhat.com/.

[19] Google, "Tesseract OCR – opensource.google.com." [Online]. Available: https://opensource.google.com/projects/tesseract.

[20] M. Kumar Audichya and J. R. Saini, "A Study to Recognize Printed Gujarati Characters Using Tesseract OCR," Int. J. Res. Appl. Sci. Eng. Technol., vol. 5, no. IX, pp. 1505–1510, 2017.

[21] M. Mathew, A. K. Singh, and C. V. Jawahar, "Multilingual OCR for Indic Scripts," in 2016 12th IAPR Workshop on Document Analysis Systems (DAS), 2016, pp. 186–191.

[22] S. M. L, A. P. J, and S. D. N, "Document Image Analysis Using Imagemagick and Tesseract-ocr," Int. Adv. Res. J. Sci. Eng. Technol., vol. 3, no. 5, pp. 108–112, 2016.

[23] N. Sankaran and C. . Jawahar, "Recognition of printed Devanagari text using BLSTM Neural Network," in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), 2013.

[24] P. Chakraborty, A. Mallik, and A. Professor, "An Open Source Tesseract based Tool for Extracting Text from Images with Application in Braille Translation for the Visually Impaired," Int. J. Comput. Appl., vol. 68, no. 16, pp. 26–32, 2013.

[25] C. Patel, A. Patel, and D. Patel, "Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study," Int. J. Comput. Appl., vol. 55, no. 10, pp. 50–56, 2012.

[26] M. A. Hasnat, R. Chowdhury, and M. Khan, "Integrating Bangla Script Recognition Support in Tesseract OCR," 2009.
[27] R. Kumar, V. Mansotra, and T. Kumar, "A first attempt to develop a lexical resource for Dogri language: Dogri WordNet using Hindi WordNet," in 2014 International Conference on Computing for Sustainable Global Development (INDIACom), 2014, pp. 575–578.
[28] M. Humayoun, R. Muhammad, A. Nawab, M. Uzair, S. Aslam, and O. Farzand, "Urdu Summary Corpus," pp. 796–800, 2014.
[29] V. Gupta and G. Singh, "Automatic Punjabi Text Extractive Summarization System," Proc. 24th Int. Conf. Comput. Linguist., vol. 2, no. December 2012, pp. 191–198, 2012.
[30] "Tesseract OCR – opensource.google.com." [Online]. Available: https://opensource.google.com/projects/tesseract.
[31] "ImageMagick: License," ImageMagick.
[32] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," J. Electron. Imaging, vol. 13, no. 1, pp. 146–165, 2004.

Ms. SonamGandotra pursed Bachelor of Computer Application from Govt. College for Women, Parade, University of Jammu, J&K in 2011 and Master of Computer Application from University of Jammu, J&K in year 2014. She is a gold medalist in MCA and has qualified NET-JRF and SET examinations. She is currently pursuing Ph.D. from Department of Computer Science & IT, Central University of Jammu, J&K. Her area of interest are Natural Language Processing and Text Mining

Dr. Bhavna Arora pursed Bachelor of Computer Science from Kurukshetra University and Post Graduated from Institute of Management Technology, Ghaziabad. She completed Ph.D. from University of Jammu, in the year 2011.She has a total work experience of 21 years in industry and academia. She is presently working as Assistant Professor in Department of Computer Science & IT, Central University of Jammu, Jammu. She is a member of IEEE, ACM, CSI, SIE, ISTE, IETE. She has published more than 28 research papers in journals and conferences of national and international repute. She has attended international conferences sponsored by DST and has also received UGC-BSR grant for project.