9. Digital Library @ IIIT Allahabad

Dr. M.D. Tiwari, UDL Project, IIIT-Allahabad

INTRODUCTION

The Indian Institute of Information Technology, Allahabad has been assigned the work of Mega Centre under DLI programme. It is working for the last more than one year. The focus of the Centre is developing specific technologies relevant to the society and to the programme. To test the validity, the Centre is also doing digitization of various books specifically in Sanskrit and Hindi. During this period, we have succeeded in developing the technologies to satisfactory level. In the present article we will focus our efforts and achievements along with certain constraints and issues faced by us. The efforts to overcome these issues and problems for developing relevant technologies will also be discussed.

CONTENT DEVELOPMENT

The focus of content development at IIIT-A has been on Hindi and Sanskrit texts since Allahabad and its neighboring regions are quite rich in libraries containing vast amount of literature in these languages. Some documents in English have also been generated. After initial surveys of the existing libraries, negotiations were initiated with several libraries namely:

- National Academy Of Sciences
- Ganga Nath Jha Sanskrit Vidyapeeth
- Allahabad Highcourt (Allahabad Law Journals)
- Hindustani Academy
- Vigyan Parishad
- Prayag Sangeet Samiti
- GB Pant Institute
- Allahabad University

Most of the libraries have cooperated with us and have started lending books / document / reports / journals for the purpose of digitization. In the following we shall discuss some of the issues that came up during the process of content generation.

Quality of documents:

Perhaps the most important issue is the quality of the documents. Since only non-copyrighted material can be put on the web servers for the general public, so most of the documents were more than sixty years old. It may be noted that the paper and print quality were not very good in most of the documents and the passage of time had resulted in further damage of the documents. Moreover, the government policy of selecting the bidder with the lowest price results quite often in choosing paper and print quality that may not be the best available. While this policy may make excellent economic sense, it creates severe problems for the process of content development. Thus, the quality of the images of the texts is usually quite low. This leads to a serious challenge in terms of cropping, cleaning and restoring the images as well as using optical character recognition systems for converting the images to text. The existing OCR systems work well for fresh documents but give poor results for these highly degraded images. In Fig. 1 below we can clearly see the type of degradation that is common. It may be noted that this is one type degradation only which is at the character level. There are other forms of degradations that we will see in later sections.

Fig. 1: A sample character 'ka'. On the left we have a well-formed character and on the right we can see a degraded character.

One can see from the above samples that no amount of image processing is going to help remove the degradation, particularly when we are talking about millions of pages. This problem has to be tackled in the character recognition process. This implies that extremely robust segmentation and recognition process has to be developed. We shall discuss in later sections the technology development corresponding to these challenges.

Copyright issues:

One of the problems that we faced was that most of the libraries are unable to supply only non-copyrighted

material. In fact, a substantial amount of the material is distinctly copyrighted. Since these documents also contain valuable information and knowledge, so we decided to digitize them along with the non-copyrighted material. However, to tackle the legal issues we have decided to approach the problem from two different directions. The first was legal and the second technical.

It was decided to approach the libraries concerned and seek their help in getting special permission from the publishers to allow the contents to be put on the web. This is turning out to be complicated process and is still in progress. The technical approach to the problem was to put a summary of the original document on the web. This, of course entails the development of good document summarizers in Hindi and Sanskrit.

Document summarizing helps in two possible ways. The first, as mentioned above, is that it allows us to get around the copyright issues. A second major advantage is that it allows us to develop very efficient search methods that will be extremely beneficial to the users. To appreciate this aspect we should recall that the digital libraries will contain millions of documents. It is not possible for a single human to read all these documents. Moreover, any single person will be interested in a small fraction of the total available documents. Therefore, we need to develop special methods that will allow the users to search for those documents that are of interest to him / her. In this context the metadata is of limited use. Quite often it contains very sketchy information about the concerned document. Typical pieces of information are name of document, name(s) of author(s), publisher, year of publication etc. While these information may help narrow down the search space, it is not sufficient. The user will be able decide the usefulness of a document only after spending a considerable time with it. On the other hand, if a good summary can be made available to the user, then a decision about the usefulness can be made quickly, thus making the whole process very user friendly.

Noting the usefulness of summaries, it was decided to develop good document summarizer. It is envisaged that the summarizer will provide syntactically and semantically correct summaries that will be tunable (i.e. the length of the summary would be a user definable quantity). Typically, very short summaries will be used with the search process and longer summaries can be generated for more in-depth studies. The work on creating a document summarizer is in progress and we shall discuss the corresponding technological developments in later sections.

Metadata:

Preparation of metadata is another major issue that has to be addressed with care. As mentioned previously, the metadata currently being used is very sketchy and carries little information. However, collecting additional information to correctly categorize the documents is a non-trivial task given the large volume of documents and non-standard categorization of themes / topics. However, availability of proper metadata is crucial for performing efficient search of the documents. We are conducting research in an effort to develop systems for automatically generating additional information about the documents. This information will be made available to the search engine so that the user can narrow down the search space efficiently and find the relevant documents. The research involves identifying the key words associated with a particular document and storing it with the other metadata of the document. This will make the key word based searches more efficient and narrow down the search space efficiently. Summaries of the documents with a high search score will be generated by the system on request by the user.

CORE TECHNOLOGY DEVELOPMENT

The earlier section contains a description of some of the issues that we are facing with respect to the process of content development. A careful analysis of the total process suggests that at least some of the difficulties can be reduced / removed by developing specific technologies. In the following we shall discuss some of the core technologies that are being developed at IIIT – Allahabad. Broadly speaking, the technologies can be categorized under three headings:

- Development of a robust OCR (with workflow) for noisy / degraded images
- Development of an efficient search engine together with a document summarizer

- Developing multi-modal user interfaces, particularly for handicapped
- Developing test bed content for access on the web

In the following subsections we shall describe our experiences and efforts in the development of each of these technologies.

1. OCR with workflow:

i) Cropper to remove the noise: An analysis of the process of digitization of documents revealed that a substantial amount of time was spent by the operators in cropping of the images, detecting and correcting skew and in removal of noise. The Scanfix software was able to take care of some of the problems but manual cropping was a pre-requisite. In order to reduce the time of operations, it was decided that an automated cropper be developed. The goal was that the cropper should be able to accept uncropped images (either single pages or complete documents stored in a single folder) and produce cropped images with minimal manual intervention. This automated cropper has been built and tested for documents in English and Devnagari scripts, for various fonts and different font sizes. In the figures below we show screen shots of the cropper. The user interface is particularly simple (since the cropper is meant to be used by the operators involved in the digitization process). The Fig. 2 demonstrates the results of cropping and in Fig. 3 we can see the effect of noise removal and de-skewing.

Fig. 2: The automated cropper

Fig. 3: De-skewing and noise removal using automated cropper

ii) OCR for degraded images: As discussed previously, the degradation of the individual characters can not be handled using the standard image processing techniques. It was essential that a robust OCR be developed for these documents. Initial efforts were made using the well known techniques of template matching and artificial neural networks. While these methods worked well for non-degraded images, the results were unacceptable for the type of images shown in Figure 1. The results of recognition were, quite often, as low as ~ 55%. To address this issue we have proposed an approach based on Gabor Feature Extraction to develop OCR for Devanagari script for such images. We adopted simple classification and recognition techniques to build a highly accurate OCR system. Experiments have been performed with a large set of characters having different types of fonts and noise variations inherent in them. We have also experimented with different classification techniques suitable with Gabor feature extraction. The initial prototype has produced very promising results and merits development into a full scale OCR system. Looking to the complexity involved in the problem following steps are required to built an efficient OCR system for very noisy images:

Efficient segmentation: This is one of the most critical requirement for the success of any OCR. In early stages of the project the segmentation was assumed to be completed and efficient. But as the project reaches the recognition stage a very high efficiency segmentation was required (at least 99%). Therefore the implemented segmentation was considered as inefficient. Thus a variety of new approaches are being tried out. By analyzing the noise present at the character and the word level, it is required that segmentation should be prepared to handle the nosiest of the noise.

Rotation and Translation invariance: Achieving rotation and translation is essential for the success of recognition of the segmented characters. A centroid based method is being tested for this purpose. This will compensate for the low efficiency in segmentation.

Integration with Multilingual Text editor: At the final step the OCR has to be integrated with a multilingual text editor with options for spell check. The OCR will be providing its output as a stream of Unicode characters with some formatting information. This output would be passed as the input in editor for final output.

Multilingual editor: A complete multilingual word processor has been developed for various Indian languages and English. The output of the editor is compatible with standard word-processors like MSWord. While the editor can operate on a stand alone basis, an API version of the editor has also been developed so that integration with other applications can be performed easily. The editor provides all the functionalities of a standard word processor and saves the files in Unicode in RTF format. Unlike many word processors, like MSWord, it is also able to save images embedded in text in an extended RTF format. In the figure below we show a screen of the editor.

Fig. 4: Screen shot of the editor while inserting images

2. Automatic search and indexing tools:

While the development of the OCR represents one dimension of technology development, the other major focus area is the development of technology for multilingual search and indexing. The specific target languages are Hindi and English in the initial phases. The major components in this area are the summarizer, annotation tool, and search engine.

Text summarization: The technique used generates an extract of the document which is a concise document description that is more revealing than the title but short enough to be absorbed in a single glance. This software can be used for rapid "relevance assessment". The user can specify the amount of summarization desired and the software generates an extract of the given document which conveys the main idea expressed in it. User can also opt for showing the "important keywords" in the document. This Document Summarizer can be used as an API by other software and also has a well developed graphical user interface. The approach involved uses pronoun and other anaphora resolution followed by statistical methods like latent semantic indexing. This approach uses only syntactic information, so further improvements using semantic analysis is under

Fig:5 Test run of document summarizer

progress. The method has been tested for English documents only and further tests on Hindi documents will be performed. In the following figure, we show the test run of the document summarizer. The left frame shows the complete document and the right frame contains the summary.

Information retrieval system: The IRS works well to identify associations in a specific corpus which is a major goal of the project. The approach is essentially based on the keywords identified in the documents. This project will be deployed on the clustered data servers in UDL project which would be logically clustered on the basis of genre (specific corpus). The future scope of this work is to include bilingual support (for English and Hindi). The approach undertaken, being independent of specific language constraints, can be easily extended for Hindi too.

3. Multi-modal interface for physically challenged:

User interface for the physically challenged people is being developed. In the initial stages we are concentrating on developing text-to-speech converters for Hindi and English. Since ready made speech corpus was available for English, so the complete TTS has been developed in this language. The corpus for Hindi is being developed currently. The other focus area is the development of interface for the people with hearing impairment. A sign language based interface is being developed for them.

Text-to-speech: The initial prototype for the Hindi TTS is ready which is able to perform the phoneme and di-phone concatenations smoothly. The audio database for Hindi TTS is being developed and is likely to involve ~ 200 hrs. of audio recording by a professional speaker. Rules for prosody are being developed and will be incorporated. TTS for English is based on the approach and data-base provided by CMU. In addition, prosody rules have already been incorporated and the system has been tested. The initial prototype is now ready, however, since the data-base has been taken from CMU, so the accent is distinctly North American. Thus, a new speech corpus is required for delivering speech in Indian accents.

Multimodal user interface for persons with hearing impairments: The text to sign language transliteration tool was implemented successfully using Java3D API. Facility to add new words and delete or update old words was provided through an easy to use user interface which even a layman can use. The speed of animation was kept to the optimum and the resource consumption has been brought down to as low as possible. The animation renders smoothly on almost all of the modern multimedia systems. However, there is still scope for modification that is going on. In Fig. 4 below we show a screen shot of the system.

Fig. 4: A screen shot showing the process for modifying an existing word

4. Developing test bed content for access on the web:

Streaming video & Video conferencing: This utility uses the RTP and SIP protocol for setting up streaming video and video conferencing facility. In addition, it provides messaging and white board services.

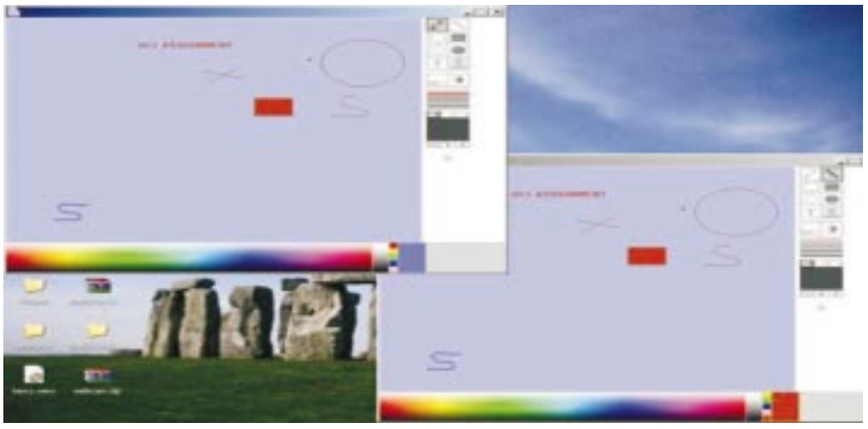

Fig. 5: a) A screen shot of the messaging and white board services.

b) A screen shot of streaming video

Annotation system: Content Management System has been developed as a part of Annotation System to maintain the available records (e-books, media files,…etc or collection of these represented as course file) in a well organized way and to maintain the users associated in the process.

CONCLUSION

In conclusion, the IIIT-Allahabad has started the work on various technologies, particularly for Hindi. These are at different stages of development. However, most of the prototypes have yielded encouraging results. While we look forward to completing these over the course of the next couple of years, we realize that some additional technologies also require our attention. Some of these are

- Hindi speech to text as part of MMUI.
- Multimedia storage, search, retrieval and display in real time for preserving other cultural heritages like music, 3D images of sculptures etc.

In addition, we also need to develop a searchable Hindi corpus. This will involve the development of Hindi NLP toolkits like part-of-speech tagger and parser.