Neural Model for Content Extraction in Multilingual Web Documents

Kolla Bhanu Prakash
Research Scholar
Sathyabama University
INDIA

M A Dorai Ranga Swamy
Professor & Dean
AVIT
INDIA

Arun Raja Raman
Professor(Retd.)
IIT MADRAS
INDIA

ABSTRACT

Neural model for multilingual web documents in Indian sub-continent is gaining prominence in day to day life. While translation and transliteration are gaining its importance on web pages, it becomes difficult for the common man to understand what the web page says about, especially when regional language is not known to the user. So, our effort here is a generic tool applied in Neural networks to overcome this problem. The model takes inputs in both English and Telugu, an Indian regional language in both printed and handwritten formats. Words having common content are chosen and neural network is used to normalize the output. A sample page from a physics textbook dealing with magnetism is taken for consideration for this paper.

General Terms

Artificial Neural Networks, Media mining.

Keywords

Media mining, Multilingual, Web communication, Neural network.

1. INTRODUCTION

Mobile networking has become a major technological innovation spreading far and wide and reaching almost every part of India. This may trigger the use of internet and web-enabled dissemination of knowledge and data worldwide. But the major problem is the language and dialect diversity even within hundred kilometers and though oral communication is possible internet and web-based communication calls for a different approach. This is particularly so in education where communication and teaching are done in English-regional language format where common words like 'cell-phones' etc are kept as they are and interwoven with regional dialect. This has paved the way for web-pages being multi-lingual and a typical web page in one of the regional language and English is shown in Fig.1.Here a telugu magazine web site gives information in both the languages. Use of similar web pages for education will considerably help the teachers and secondary and higher class students. But to generate these web pages one has to know the overall content of the subject to be rewritten in a multilingual format for easy understanding. The focus of the present paper is in this direction and a neural model is generated to assess the content of a web page.

Fig. 1. Typical multi-lingual web page combining English and regional language Telugu

2. FEATURES OF INDIAN TEXTS

Data mining has progressed by leaps and bounds in the last decade and has become a powerful tool with number of software backups for translating documents in different languages and in different forms. These approaches which rely on the structural features of the documents may not be directly applicable to multi-lingual documents in the Indian context mainly due to varied forms of regional language texts and the syntax formats in some cases may also be different. A letter in Indian regional language may be a combination of three different characters making it difficult to apply directly data mining approaches. Hence a media mining approach using pattern recognition at pixelmaps [1-4] is needed. Fig.2 gives the English text 'Bar' and 'andhra' and corresponding regional language text written in regional language, Telugu; respectively. These are the same words used for tested input.

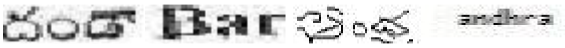


Fig.2 Features of Indian multi-lingual text

3. PREPROCESSING AND FEATURE VECTORS

Pixelmaps of selected words from web pages are taken in the form of jpg files and normalized to a predefined size of nxp by nyp. From these pixelmaps the features of the pixelmaps are extracted as vectors with twenty elements and these are used in the neural model as inputs. Different sets were used in the neural model and by varying the weighted matrices according to content relation; training of the inputs is done to achieve the target output as binary one with content related being one and not related being zero. Three sets of pixelmaps representing content words are used in the basic model with ‘magnet’, ‘iron’ and ‘filings’ etc with corresponding regional ones.

4. TYPES OF DOCUMENTS

Another dimension in this procedure is the variation in the type of documents from which input pixelmaps are obtained. Here computer-generated texts-CT-.printed texts-PT and handwritten texts –HT- are used for the documents. While CT pixelmaps will be crisp, both PT and HT will have non-crisp and noisy inputs depending on the nature of the parent document. Fig.3 shows for magnet all three forms.

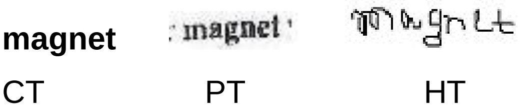


Fig.3 Document variations

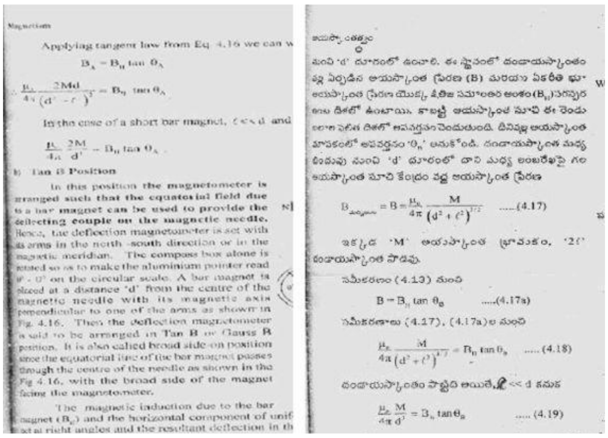
The feature vectors were developed using the same method for all the cases so that training the sets will be able to account for vagaries in the inputted documents.

5. METHODOLOGY

Standard neural network approach is coded with weights and transfer functions as parameter variations. Training with basic data inputs was done with a) pixelmaps from printed texts and later with b) computer-generated texts. Since computer-generated texts showed closeness with normalized values within 4%, they were used as the basis for future testing. Since the focus of model is to arrive at the content, a threshold was used for developing the target with 15% variations.

6. ILLUSTRATIVE EXAMPLE

A typical textbook page on magnetism is shown in Fig.4 with both English and regional Telugu text. Here one can observe that the printed document contains text fonts different from a regular word processor as these were done from a type-written text. Fig.4 (a) gives the English text page related to magnetism and the corresponding one in regional language Telugu in Fig.4 (b). It may be seen that equations are retained in English while the text is in Telugu making a multi-lingual presentation.



Magnetism

Applying tangent law from Eq. 4.16 we can w

$$B_A = B_H \tan\theta_A$$

$$\therefore \frac{\mu_0}{4\pi}\frac{2Md}{(d^2-\ell^2)^2} = B_H \tan\theta_A$$

In the case of a short bar magnet, ℓ << d and

$$\frac{\mu_0}{4\pi}\frac{2M}{d^3} = B_H \tan\theta_A$$

b) Tan B Position

In this position the magnetometer is arranged such that the equatorial field due to a bar magnet can be used to provide the deflecting couple on the magnetic needle. Hence, the deflection magnetometer is set with its arms in the north -south direction or in the magnetic meridian. The compass box alone is rotated so as to make the aluminium pointer read 0° - 0° on the circular scale. A bar magnet is placed at a distance 'd' from the centre of the magnetic needle with its magnetic axis perpendicular to one of the arms as shown in Fig. 4.16. Then the deflection magnetometer is said to be arranged in Tan B or Gauss B position. It is also called broad side-on position since the equatorial line of the bar magnet passes through the centre of the needle as shown in the Fig 4.16, with the broad side of the magnet facing the magnetometer.

The magnetic induction due to the bar magnet (B_B) and the horizontal component of unif

at right angles and the resultant deflection in th

అయస్కాంతత్వం

నుంచి 'd' దూరంలో ఉంచాలి. ఈ స్థానంలో దండాయస్కాంతం వల్ల ఏర్పడిన అయస్కాంత ప్రేరణ (B) మరియు ఏకరీతి భూ అయస్కాంత ప్రేరణ యొక్క క్షితిజ సమాంతర అంశం (B_H) పరస్పర లంబ దిశలో ఉంటాయి. కాబట్టి అయస్కాంత సూచి ఈ రెండు బలాల ఫలిత దిశలో అపవర్తనం చెందుతుంది. దీనివల్ల అయస్కాంత మాపకంలో అపవర్తనం 'θ_B' అనుకోండి. దండాయస్కాంత మధ్య బిందువు నుంచి 'd' దూరంలో దాని మధ్య లంబరేఖపై గల అయస్కాంత సూచి కేంద్రం వద్ద అయస్కాంత ప్రేరణ

$$B_{\text{మధ్యమ}} = B = \frac{\mu_0}{4\pi}\frac{M}{(d^2+\ell^2)^{3/2}} \quad \text{......(4.17)}$$

ఇక్కడ 'M' అయస్కాంత భ్రామకం, '2ℓ' దండాయస్కాంత పొడవు.

సమీకరణం (4.13) నుంచి

$$B = B_H \tan\theta_B \quad \text{.....(4.17a)}$$

సమీకరణాలు (4.17), (4.17a)ల నుంచి

$$\frac{\mu_0}{4\pi}\frac{M}{(d^2+\ell^2)^{3/2}} = B_H \tan\theta_B \quad \text{.....(4.18)}$$

దండాయస్కాంతం పొట్టిది అయితే, ℓ << d కనుక

$$\frac{\mu_0}{4\pi}\frac{M}{d^3} = B_H \tan\theta_B \quad \text{.....(4.19)}$$

a) Page from English text **b) Page from Telugu text**

Fig.4 Text book page in two languages.

Choosing selected words from this document as input pixelmaps, variation in target with respect to the word ‘magnet’ in English and telugu is shown in Fig.5. Here all words chosen from printed texts-PT- have a variation within 22% and in each set the second and fourth feature perform very well close to 1.

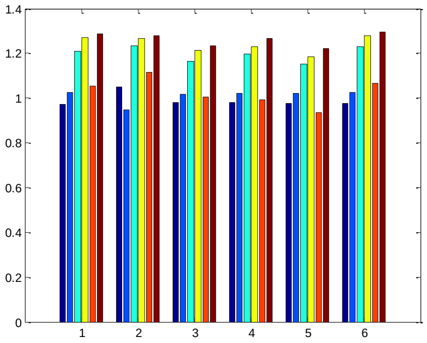


Fig.5 Histogram variations in printed texts

Now the same texts were generated with computer word processor and using these as basic inputs, the variations were plotted as in Fig.6 and here all the variations fall within 10% except for the fourth feature.

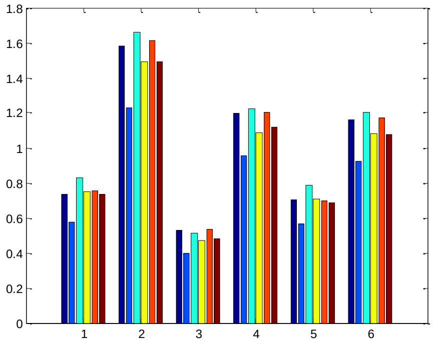
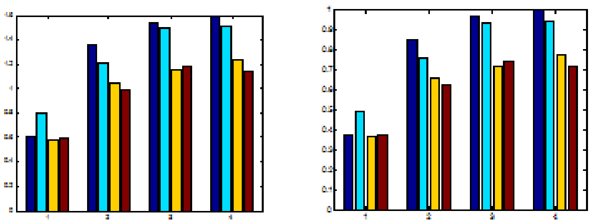


Fig. 6 Histogram variations with computer generated texts

Now the basic input/output pattern for computer generated texts can be used for normalization as certainty of content is assured. With this as the premise the neural model is used to predict the content of texts from different web-based documents. Five words were chosen with first three of them not related to magnetism and fourth and fifth related to magnetism. Typical results for these text pixelmaps are shown in Fig.7, for prediction with printed and computer-generated training. In Fig.7(a) the closeness of the last two results indicate content as magnetism and it is getting better with computer generated texts as in Fig.7(b).



a) Based on printed text b) Based on Computer generated Fig.7 Content prediction

7. CONCLUSIONS

Web pages in multi-lingual format [7-12] are getting prominence in Indian education scene and the paper proposes a method using ANN to assess the content irrespective of the language of the input texts. It is proposed to extend this to images and other media types as well. Work is carried out with respect to statistical interpretation and ANN.[5,6] In order to overcome some discrepancies in these two methods, Our future work is focused to combine the two methods and work will be carried out with different sample inputs.

8. REFERENCES

[1] Rafael C. Gonzalez, Richard E. Woods, Steven L. Eddins, “Digital image processing using matlab”,2002.

[2] Renu dhir, “Feature extraction and classification for bilingual script (Gurumukhi and Roman)”, April 2007.

[3] Bing Zhao, Stephen Vogel, “Adaptive parallel sentences mining from web bilingual news collection”, 2002.

[4] Y. Li, C.-C. J. Kuo and X. Wan, Introduction to content-based image retrieval — Overview of key techniques, in Image Databases: Search and Retrieval of Digital Imagery,eds. V. Castelli and L. D. Bergman (John Wiley, New York, 2002), pp. 261–284.

[5] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, “Statistical Interpretation for Mining Hybrid Regional Web Documents”, ICIP 2012, CCIS 292, pp.503–512, 2012 © Springer-Verlag Berlin Heidelberg 2012.

[6] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, “ANN for Multilingual Regional Web Communication”, ICONIP 2012, Part V, LNCS 7667, pp. 473–478, 2012, © Springer-Verlag Berlin Heidelberg 2012.

[7] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, “Performance of Content Based Mining Approach for Multi-lingual Textual Data”, International Journal of Modern Engineering Research, Vol.1, Issue1, Sep-Oct 2011, pp-146-150.

[8] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, Content Extraction with Web Pages having Hand-Written Texts” (NCEVENT 2011) Sathyabama University, Chennai.

[9] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, “Text Studies Towards Multi-lingual Content Mining for Web Communication” (TISC2010), Sathyabama University, Chennai.

[10] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, “Content Extraction for Multi-lingual Web documents”, CIT Journal of Research Volume 1, Issue 3, nov 2010, pp.93-101, Chhattisgarh Institute Of Technology, Rajnandgaon.

[11] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, A Neuron Model for Documents Containing Multilingual Indian Texts (ICCCT 2010), Allahabad.

[12] Kolla Bhanu Prakash, M.A.Dorai Ranga Swamy, Arun Raja Raman, “Feature extraction for content mining in multi-lingual documents” (NCICN 2010), Sathyabama University, Chennai.