Adaptive Semantics Web-based Approach for E-education using the Static Annotated Corpus

Prakash Choudhary[1] and Neeta Nain[2]

[1]Department of Computer Science and Engineering
National Institute of Technology Manipur
Imphal, India-795001
Email: choudharyprakash@nitmanipur.ac.in

[2]Department of Computer Science and Engineering
Malaviya National Institute of Technology Jaipur
Rajasthan, India-302017
Email: neetanain@yahoo.com

ABSTRACT

In recent era popularity of E-education has been increased. Task of selecting an appropriate content for right education become a real challenge. This paper outline a semantic web based approach for E-learning using the static corpus platform. Adaptive hypermedia semantics technology using the available content become a one of the feasible solution for E-learning. In this paper, we describe a structure that use the semantic technology for exploring the database information by allowing intelligent navigation, personalize of information, retrieval of information by searching and querying. We also describe in detail about the web-based annotation and visualization of content and way to explore the personalized information in both printed and handwritten format on same viewport. The purpose tool also supports the functionality of finding, manipulation and organization of the new content.

KEYWORDS

Corpus-based language learning, Corpus integration, Web based interface, ICT, Corpus education.

1 INTRODUCTION

The current web-based design is performing the interpretation between the systems and humans. In digital trends of Information and Communication Technology (ICT), the medium of education gradually being a part of on-line education or system based education instead of documents based education. The general awareness and prominence of digital communication technology increasing day by day between the peoples. Today’s education sources mainly depend on the computer, mobile, and the internet. Semantic web-based technology is utilizing these sources fruitfully. The semantic-based corpus structure is platform independent and can be accessed on any internet connected devices. Semantic web-based technology is useful for build a system that can support both web plus data or sorting of data from the database. Semantic web provides a useful way to develop the system that can support the interpretation of various computer and mobile devices over the network.

The evolution of corpus-based linguistic and language learning has gained prominence over the last four decades. The strength of corpus can be defined as the large collection of natural language texts, where the genre of text may be handwritten, spoken and printed to characterize all hypothesis about the language. The evolution of Corpus linguistic is started from since 1980. From 1990’s researchers increased the interest in applying the findings of corpus-based research to language pedagogy. The upsurge interest of researcher’s community towards the teaching and language corpus can be seen from the following literature [1]-[9]. Wichmann [10] observed a direct or indirect convergence between corpora and teaching.

As from the available sources, it has noted that during the development of the corpus main focus was given to the utilization of corpus in the field of linguistic or natural language processing domain. Therefore, the aim of the development of a corpus is surrounding around the document specific instead of the corpus structure. Lack of standard tool to collaboration between language teaching and corpus can be considered as one of the reasons for slower development of corpus based education. Considering these issues, we build a structure where the user can explore the both handwritten document and associated the computer readable information on the same platform simultaneously. Because so far there does not exist a structure where we can explore both the handwritten image and associated Information simultaneously.

Here, we deign a structure where the material of printed, Unicode or handwritten text can explore on a single system. The Same viewport displays the electronic text along with the handwritten image of that text. For learning a language handwritten part is also playing an important role because the style of writing for same text may be different based on writers. In some cases, the characters take a different shape during the writing. For example in the language Urdu writers can join the characters at the time of writing which looks different from the printed texts also called ligatures.

The structure we have developed for building a corpus considers a paragraph or full-length sentences. Where user can explore the detail of each paragraph in both top-down or bottom-up approach. From the word we can reach at the sentence to find out the uses and meaning of words in this sentence and the image of handwritten word display at side panel. An annotated information has associated with each words such as: synonyms, antonyms grammatical information (Noun, Pronoun, Verb, etc.) of the word to build a dictionary. Structure also allow user to personalize navigation to explore the information and user able to download the desired result in Unicode, handwritten or XML format.

The paper is organized as, Section II, background study of the corpus to explore the direct or indirect contribution of different types of corpora in language teaching and learning. Section III Describe the process of personalizing of the corpus for making it suitable for E-learning. Section IV discusses the structure of web based system. Section V describes the mapping of coordinates with structure for visualization of handwritten text region. At last in section VI conclusions are presented.

2 BACKGROUND STUDY

Large scale computerized collection of realistic texts samples are stored in a systematic way to use in the vast field of linguistic research for various purposes. Corpora are significant for the lexicographers in several ways. Machine readable nature of corpus giving the greatest advantage in the lexicography. That allows language learners to extract all authentic information about the lexical item with typical real life examples. Corpora could be used extensively to provide more accurate descriptions about the use of language, language testing [14] and design syllabuses including teaching materials [11][13]. Coniam [15] shown the demonstration about the development of test materials and optimize the test procedures with the processing of an annotated corpus as a training database.

Linguistic finding out driven from the corpus is divided into two types based on extraction process and their results [12]. Corpus-driven linguistic is a bottom-up approach where finding is validated by corpus evidence and meaning of words has to be negotiated during the annotation. Corpus-based linguistic is a top-down approach where meaning is first made and corpus data is used for testing and improve the prior knowledge [30][31]. In this case, the finding is validated by grammatical features such as antonyms, synonyms, and tagging. Types of the corpus are mainly divided in the following terms:

- Spoken Vs. Written
- Static Vs. Dynamic
- Plain Vs. Annotated
- Parallel or Multilingual

These typologies of the corpus are defined the terms relevance to corpus. For example, plain corpora are raw collection of text while annotated corpora associate the mark-up information with plain texts. It can be an analysis that none of the corpora is superior to the rest because each corpus has different utilization based on the nature of corpus, the parallel corpus is useful for translation and making bilingual dictionaries.

Corpus linguistic found useful to reveal hidden facts of language such as grammatical, morphological, syntactically and semantic [32]. In 1995, Susan, argues the first time to use the computer-stored English corpus for grammar awareness [26]. Based on the argument they observed that use of corpus provides an excellent data for doing grammar. General corpora like British National Corpus (BNC) [25] is a static corpus having an extensive collection of spoken and written text sample around 100 million words. These large corpora are providing many features which is helpful for research as well as learning The English language, and it is also divided into subcorpus for analysis on particular domain. Open American National Corpus (OANC) [27] is another example of the corpus that focus on the American English. COCA (Corpus of Contemporary American English) [28] is a collection of texts from spoken and printed. The structure of COCA corpus allows the user to search the words, phrases, and grammatical constructions.

Some books and work explore the techniques and way to the utilization of corpus in the field of teaching and learning using the corpus [1]-[10]. However, application of corpus linguistics in the area of education is not extent as it expected. One reason is that teachers and learners do not adopt the ideology and technique of corpora based education. Second, most of the corpus are built to fulfill research requirement during the development most concern is surrounding on document specific instead of the corpus structure.

3 STRUCTURE OF CALAM

In this paper, we have designed a closed Corpus adaptive system. The structure can operate on the documentation that has been developed using this structure. Where the relation between the documentation and systems are defined at the time of systems design. But the design of the structure is flexible, the desired document can be annotated as per requirement.

Corpus development focus was independent of the structure, which make it difficult to available the corpus for teaching and learning the language. In our work, we designed a structure where a developer can build a corpus to display handwritten text and corresponding transcription on the same screen. The name of structure to develop and annotate corpus is CALAM (Cursive and Language adaptive methodology). The salient features of the structures is as following:

- Structure providing the following functionality to develop a standard corpus in systematic way for various purpose. The corresponding brackets shows the availability of function according to users:
    - Insert(developer)
    - Delete(developer)
    - XML Markup Generation(user)
    - Select color Scheme (theme layout)(user)
    - Update(Edit Data)(developer)
    - Search(to search the information from database using keyword, ID)(user)
    - A to Z(word list)(user)
- Nature of structure is platform independent it can be access with any devices.
- Structure currently supporting 3 languages, but it can be extend to support for other languages.
- For balancing among the contents/material domain of data-collection is distributed among

6 categories which is further divided into subcategories.

- Handwritten data collection form is designed in a specific way to collect large amount of printed and handwritten texts together.
- For each handwritten paragraph, sentences, lines and word, transcription electronic text display in side panel.
- Information can be retrieve from the corpora, either in form of image, cluster list, paragraph, and word list.
- Calculate the frequency of words occurrence with their uses.
- XML file format in a hierarchical manner having complete information for research findings.
- Enable user to explore the list of vocabulary and grammar structures.

4 PERSONALIZE FOR E-LEARNING

From the educational perspective, the corpus is an enormous collection of documents material that covers various topics. Each user has different requirement related to contents. So it is desirable to personalize the content based on need.

The structure provides support for extraction of required data from texts, processing and interpretation of the output according to user needs. Software has a number of advantages from a user perspective: it produces the list of words, count the frequency and occurrences of individual search items, allow user to present and organization of data in a way that facilitates the identification of patterns.

Figure 1 shows the block architecture for retrieval and personalizes of information. The content can be explored on any internet connected devices computer, or mobile and database will be stored on the server side. A layer of personalizing of information is work for filtering and extraction the desired data from the corpus, deliver presentation of requested content according to needs. A web based structure CALAM having potential to use the corpus for teaching and learning the purpose and facilitates it with the following functionality:

- Make the corpus relevant for teacher and learners.
- Design corpus-aided activity to retrieval information easily.
- Enrich in language knowledge and variations.
- Able to store and explore natural collection of language patterns instead of abstract explanations, showing realistic in use.
- Revealing the trends of historical changes in language.

4.1 Term-based content modeling and filtering

In [22] [23], describe a web based approach to the annotating corpus to support e-learning. In our approach of annotation, an automatic indexing of structured content is associated with the programming code that will be applied for both fetching and extraction of data. Entry of new document added and modelled manually based on the content category while the requested information filtering and retrieval can be done automatically based on same content terms. Annotation could help learners in highlighting the most significant part of a text, they can get more ideas about the selected texts parts.

4.2 Intelligent navigation and Information extraction

We designed Graphical User Interface (GUI), a web application called CALAM that enables the user to select data from the corpus and provide the facility for the developer to annotate the corpus by indexing the content automatically in the database. Reouf[24] also purposed a system to integrate the web based corpus linguistic.

As a result, the user can select among several of options for display the extraction information, or can download the selected portions of the corpus and annotations in their original scanned

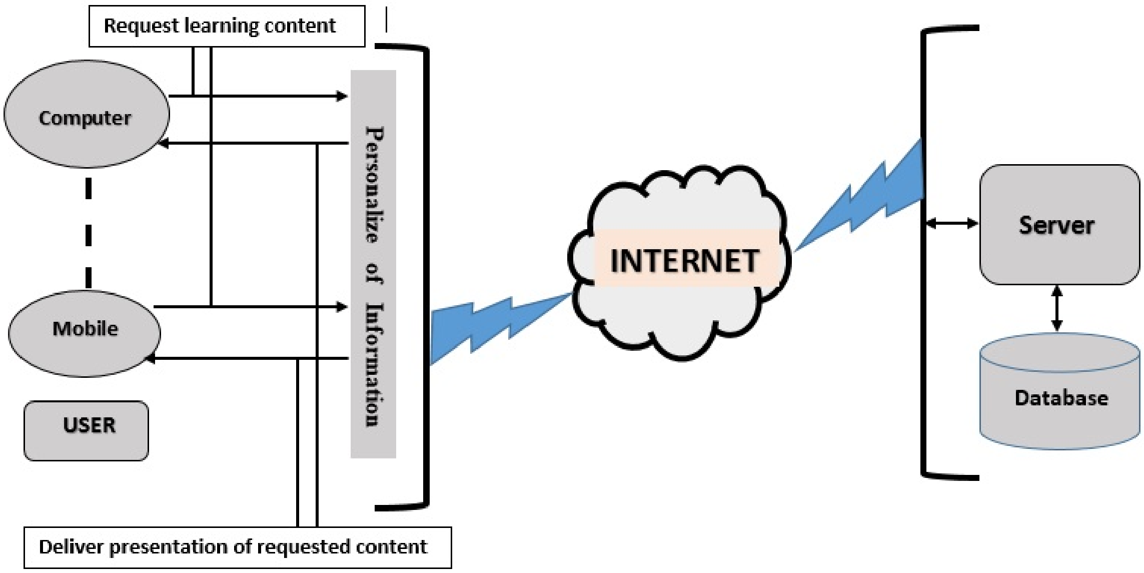
Request learning content
Computer
Mobile
USER
Personalize of Information
INTERNET
Server
Database
Deliver presentation of requested content

Fig. 1: Block level architecture of corpus personalization.

handwritten format, Unicode or XML format. The delivery of requested content is processed by merging the annotations and rendering them in the desired output format, then resulted output is available for download. Thus, the developer can create a customized corpus with data and annotations of their choice, and the result of requested content can be delivered in the most convenient way.

4.3 Semantic tagging

Strongest features of the corpus are to build dictionaries with realistic text samples. In 1987, Collins developed a first fully corpus-based dictionary Cobuild[16] for The English language. After the Cobuild dictionary, corpus was frequently used to build the dictionaries for the modern languages such as: [17] Spanish, [18] German, [19] Portuguese, [20] French and [21] Chinese.

Tagging is not only the information that can be added to the corpus while it makes corpus beneficial for the non-specialist language learner. An annotated information has associated with each word such as: synonyms, antonyms grammatical information (Noun, Pronoun, Verb, etc.) of the word. Functionality of A to Z, display all the list of words in alphabetical order and the total number of words.

5 MAPPING BASED ON COORDINATES INFORMATION

For the better visualization of handwritten text and corresponding text a mapping has been done in the structure. Mapping of structure is done at four level: image, line, word and character for depth information of materials.

5.1 Image contents

The information associate with image is as following:

- Text(handwritten texts of image)
- Number of lines(handwritten lines in image)
- Information added to available corpus for research purpose.

5.2 Lines spotting

In next step when user select particular line from the side panel. The line automatically crop from image and display along with the line image and transcription information.

- ID: An auto-indexing number generated automatically based on the image
- Text(handwritten text of line)
- Pixel information of textual region is stored in database for highlight the handwritten text for better visualization.
- Number of Words in Line

Based on selected region the structure automatically fetch the coordinates of line pixels such as upper X, lower X, upper Y, lower Y and calculate the Width and Height from X, Y coordinate. When cursor move on line ID a rectangle box will show around that line in side panel. This rectangle box will generate according to our coordinates of X, Y height and weight information which we taken during the crop of line. When we select the line the crop image of line will display along with the information.

5.3 Words spotting

In continuation of line user can select particular word from the side panel. The word automatically crop from image and display along with word information.

- ID: which is automatically generated based on the line ID
- Pixels information of textual region
- Text(handwritten text of word)
- Number of characters in Line
- Synonyms
- Antonyms
- TAG

Same process of line it will follow for the words. At the time of exploring information user can move back to line for finding the uses of word in sentence. This level of navigation explore the words with synonyms, antonyms together.

5.4 Characters spotting

To categorize the word and explore the combination of characters a rectangle box will show around that character in side panel on image. This rectangle box will generate according to our coordinates of X, Y height and weight information.

6 CONCLUSION

In this work, we tried to bring both handwritten and printed text on the same platform. The material of handwritten documents and corresponding transcription text useful for the learner to explore the style of writing for E-education. The aim of work described in this paper is to design a structure where a corpus can be built for the various purpose including the education.

Structure allows users to design a customized corpus, including handwritten document, annotations, transcription text of their choice and they can download the desired result in a format of their choosing. The web application, together with additional tools encoded data and annotations for various purpose NLP systems, build the dictionary, and vocabulary of antonyms and synonyms, frequency of words and uses of the word in sentences. The structure provides support for the selection and distribution of personalized e-learning content (e.g. graphics and text) to e-learners that will best suit for user’s interests and goals, meet their formatting preferences.

7 REFERENCES

1. Aston, G. (1995). ”Corpora in language pedagogy: matching theory and practice” Principle and Practice in Applied Linguistics(Oxford University Press), pages 257–270, 1995.
2. Partington, A. (1998). ”Patterns and Meanings: Using Corpora for English Language Research and Teaching”. Amsterdam: John Benjamins.
3. Aston , G., Bernardini, S. and Stewart, D. (eds.) (2004). ”Corpora and Language Learners”, Amsterdam: John Benjamins.
4. Burnard, L. and McEnery, A. (eds.) (2000). ”Rethinking Language Pedagogy from a Corpus Perspective”, Frankfurt: Peter Lang.
5. Kettemann, B. (1996). ”Concordancing in English Language Teaching” In Proceedings of Teaching and Language Corpora, pages 4–16, Lancaster University.
6. Meunier, F. (2002). ”The pedagogical value of native and learner corpora in EFL grammar teaching” In Second Language Acquisition and Foreign Language Teaching, pages 119-142, Philadelphia: John Benjamins.

7. Upton, T. and Connor, U. (2001). ”Using computerized corpus analysis to investigate the text-linguistic discourse move of a genre”, English for Specific Purposes, pages 313–329, 2001.
8. Tan, M. (2002). ”Corpus Studies in Language Education”, IELE Press, Bangkok.
9. Braun, S. (2007). ”Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora”, ReCALL, pages 307–328, 2007.
10. Leech, G. (1997). ”Teaching and language corpora: a convergence” in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) In
11. Conferences of Teaching and Language Corpora, pages 1–23, London, 1997.
12. Osborne, J. (2001). ”Integrating corpora into a language-learning syllabus” in B. Lewandowska-Tomaszczyk (ed.) PALC 2001: Practical Applications in Language Corpora, pages 479–492, Frankfurt, 2001
13. Osborne, J. (2002). ”Top-down and bottom-up approaches to corpora in language teaching” in U. Connor and T. Upton (eds.) Applied Corpus Linguistics: A Multidimensional Perspective, pages, 251–265, Amsterdam, 2002.
14. Mindt, D. (1996). ”English corpus linguistics and the foreign language teaching syllabus” in J. Thomas and M. Short (eds.) Using Corpora for Language Research, pages 232–247, Harlow, 1996.
15. Alderson, C. (1996). ”Do corpora have a role in language assessment?” in J. Thomas and M. Short (eds.) Using Corpora for Language Research, pages 248–259, London, 1996.
16. Coniam, D. (1997). ”A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests”, CALICO Journal, pages 15–33, 1997.
17. Harper Collins (1995) Collins COBUILD English Dictionary (2nd edition). London: Collins CoBUILD.
18. Davies, M. (2005) A Frequency Dictionary of Spanish. London: Routledge.
19. Jones, R. and Tschirner, E. (2005) A Frequency Dictionary of German, London: Routledge.
20. [19] Davies, M. and de Oliveira Preto-Bay, A. (2007) A Frequency Dictionary of Portuguese. London: Routledge.
21. Lonsdale, D. and Bras, Y. (2009) A Frequency Dictionary of French. London: Routledge.
22. Xiao, R., Rayson, P. and McEnery, T. (2009) A Frequency Dictionary of Mandarin Chinese. London: Routledge.
23. Bonifazi F., Levialdi S., Rizzo P., Trinchese R. (2002). ”UCA: A webbased annotation tool supporting e-learning”, Proceedings of the Working Conference on Advanced Visual Interfaces pages 123–128, New York (NY), 2002.
24. Brusilovsky P. (1997). ”Efficient techniques for Adaptive Hypermedia”, In C. Nicholas and J. Mayfield (eds.), Lecture Notes in Computer Science, Vol. 1326, pages 12–30, Berlin, 1997.
25. Renouf, A., A. Kehoe and J. Banerjee (2007). ”WebCorp: an integrated system for web search”. In M. Hundt, N. Nesselhauf, and C. Biewer (eds.), Corpus Linguistics and the Web, pages 47–68, Amsterdam and New York, 2007.
26. Hoffmann, S., S. Evert, N. Smith, D. Lee and Y. Berglund Prytz (2008). ”Corpus Linguistics with BNCweb-a Practical Guide”, Frankfurt am Mein: Peter Lang, 2008
27. Susan, Hunston (1995). ”Grammar in teacher education: The role of a corpus”, Journal of Language Awareness, Vol 4, Number 1, pages 15–31, 1995.
28. http://www.anc.org/
29. http://corpus.byu.edu/coca/
30. Prakash Choudhary and Neeta Nain, "A Structure for Annotation and Ground-Truthing of Urdu Handwritten Text Image Corpus" 7th International Conference on Corpus Linguistic, CILC2015, Proceeding of Elsevier, University of Valladolid, Spain, pages 84-88, 2014.
31. Prakash Choudhary and Neeta Nain, "CALAM: Linguistic Structure to Annotate Handwritten Text Image Corpus" International Conference on Computational Intelligence in Data Mining, Springer, Orissa, pages 449-460, 2014.
32. Prakash Choudhary and Neeta Nain ,"An Annotated Urdu Corpus of Handwritten Text Image and Benchmarking of Corpus" 37th IEEE International Convention, MIPRO2014, Conference on Intelligent Systems by IEEE at Croatia, pages 1409-1412, 2014.