Autonomous Educational Testing System Using Unsupervised Feature Learning.

Anwaya Aras[1], Shree Ranjani[2], Jannat Talwar[3], Dr. Mangesh Bedekar[4]

[1,2,3,4]Department of CS&IS, BITS Pilani K K Birla Goa Campus,
Zuarinagar, Goa, PIN-403726, India.

Abstract

With the increase of ubiquitous data all over the internet, intelligent classroom systems that integrate traditional learning techniques with modern e-learning tools have become quite popular and necessary today. Although a substantial amount of work has been done in the field of e-learning, specifically in automation of objective question and answer evaluation, personalized learning, adaptive evaluation systems, the field of qualitative analysis of a student's subjective paragraph answers remains unexplored to a large extent.

The traditional board, chalk, talk based classroom scenario involves a teacher setting question papers based on the concepts taught, checks the answers written by students manually and thus evaluates the students' performance. However, setting question papers remains a time consuming process with the teacher having to bother about question quality, level of difficulty and redundancy. In addition the process of manually correcting students' answers is a cumbersome and tedious task especially where the class size is large.

In this paper, we put forth the design, analysis and implementation details along with some experimental outputs to build a system that integrates all the above mentioned tasks with minimal teacher involvement that not only automates the traditional classroom scenario but also overcomes its inherent shortcomings and fallacies.

Keywords:

Machine learning, Natural language processing, Intelligent tutoring, Pedagogy Enhancement E-learning techniques, Document Object Model Parser, Information Retrieval, Web data mining.

1. Introduction

Our system broadly handles two major tasks: generating questions and evaluation of subjective answers. The system extracts questions from a webpage on the Internet using web data mining techniques based on the topic chosen by the teacher. Since the internet is a varied and colossal repository of information, it is necessary to categorize questions based on their relevance to the topic chosen by the teacher. This is done using classification algorithms under machine learning and the teacher is only required to provide the initial training set. The question database is further updated with an additional entry for difficulty level per question. The students can now take the test as per their convenience to test their understanding of the subject.

Once the student starts the test, the system simultaneously evaluates his subjective answers in real time. Based on the student's performance, the system increases or decreases the difficulty level of his questions. Subjective answer evaluation is done using natural language processing techniques where the system determines the percentage correctness of a students' answer by comparing it with a standard answer from the database. This method of checking ensures an unbiased and efficient evaluation of textual paragraphed answers.

Our system effectively analyses the level of comprehension of a particular topic by a student. Automation of the traditional classroom scenario is accomplished by making optimal usage of web resources and thus we have a robust, fully functional intelligent classroom scenario in place.

2. State of the Art

In structuring a quiz it is important to work with a theoretical framework of learning. Assessment processes are now in the limelight, with increasing emphasis placed not on testing discrete skills or on measuring what people know, but on fostering learning and transfer of knowledge. The traditional approach to assessment is largely a form of objective testing which tends to value students' capacity to memorize facts and then recall them during a test situation. Typically, objective quiz questions are considered limited in their capacity to assess cognitive skills of a student. These types of questions promote guessing and do not reveal the level of thinking of the student. They often fail to provide a learning experience for students as they only deal with factual recall.

Subjective questions can be used to test students' ability to synthesize and apply concepts. This paper introduces a system for subjective quiz generation and automatic correction and analysis of answers. Using this system, a teacher can make a deeper evaluation on the students' performance, instead of only checking his/her marks. In addition, this system fosters self-regulated learning as students can give the tests at a time and place convenient to them.

I. Proposed Architecture:

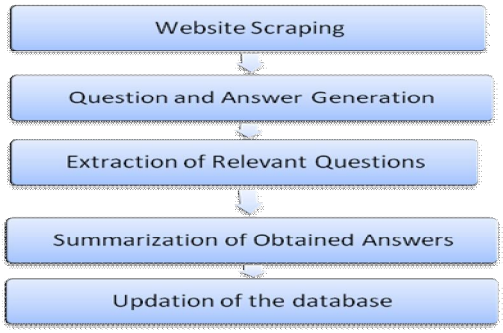


Fig1. Methodology-Teacher Interface

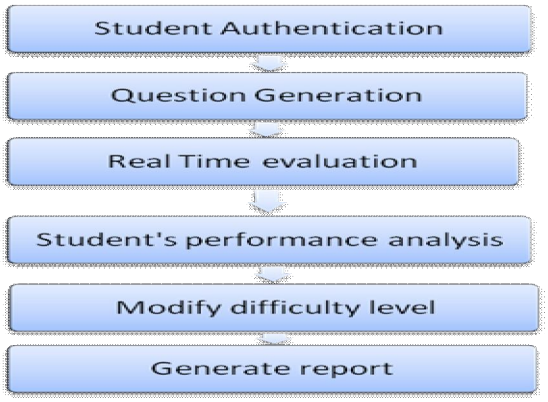


Fig2. Methodology-Student Interface

III.A. Question Generation

Generating subjective questions is the basic requirement of this system. This task has been accomplished using web data mining. Web mining adopts data mining techniques to automatically discover and retrieve information from web documents. A web data extraction system usually interacts with a web source and extracts data stored in it: for instance, if the source is a HTML web page, the extracted information could consist of elements in the page as well as the full-text of the page itself. Eventually, extracted data might be post-processed, converted in the most convenient structured format and stored for further usage [1].

Web data extraction allows to efficiently collect this information with a limited human effort. The role of the teacher in subjective question generation is limited to only providing the url of the webpage which contains the required questions and answers. All the remaining work is handled by the web data extraction system. The design and implementation of this system has been considered from different perspectives and it leverages on scientific tools coming from various disciplines like Machine Learning, Logic and Natural Language Processing. As this system involves web data mining, many factors were taken into account; some of which were independent of the specific application domain in which we planned to perform web data mining. Other factors, instead, heavily depended on the particular features of our application domain: as a consequence, some technological solutions were not suitable for this system. Some approaches focused on static HTML web pages and used the tags composing a page along with their hierarchical organization to extract information.

i. Related Work.

There is a large body of related work in data extraction and information retrieval that attempts to solve similar problems using various other techniques. Although much research has been done on data extraction, it is still a relatively new field. Laender [2], presented a survey that covers a rigorous taxonomy to classify and analyze web data mining. They introduced a set of criteria and a qualitative analysis of various Web Mining tools**.** Kushmerick [3] tracked a profile of finite-state approaches to the Web Data Mining problem. Web Data Mining techniques derived from Natural Language Processing and Hidden Markov Models have also been discussed in the past. [3]

ii.Our approach

The url of the webpage and the specific topic are taken as input from the teacher. The questions along with their respective answers are then scraped from the url and shown to the teacher for confirmation. Here we were faced with a choice. We could either use Regular expression matching or HTML DOM parsing[4] to extract the questions. Although regular expression matching was an easier option, we decided to go with the latter since it is less processor intensive and has a lesser chance of failing if the layout of the page changes slightly. The representation of a Web page by using a labeled ordered rooted tree is usually referred as DOM (Document Object Model). The general idea behind the Document Object Model is that HTML Web pages are represented by means of plain text, which contains HTML tags, i.e., particular keywords defined in the mark-up language that can be interpreted by the browser to represent the elements specific of a Web page (e.g., hyper-links, buttons, images and so forth). HTML tags may be nested one into another, forming a hierarchical structure. This hierarchy is captured in the DOM by the document tree, whose nodes represent HTML tags [5]. By parsing the webpage into a DOM tree, more control can be achieved while extracting content. This concept was used to implement the extraction of questions by this system. After confirmation, the question-and-answer sets are added to the database of questions already present. If the topic is new, it is also added to the table of topics in the database.

III.B.Text categorization

As the data obtained after extraction from internet may be varied and not be typically relevant, categorizing into topics is essential. The goal of text categorisation in this context is categorising the question according to the topic chosen.Using machine learning the objective is to learn classifiers from examples which do the categorisation automatically.The categorisation in this case boils down to binary classification as to whether the question belongs to the topic or not.
Traditionally, 5 major approaches have been used to solve this problem of text categorisation namely Naive Bayes algorithm,Support vector machines,Artificial neural networks,k-Nearest neighbour classifier,C4.5 decision tree learner.SVMs are the best conventional method in comparison with others excluding Artificial Neural Networks.[6]Support vector machines with its ability to handle high dimensional input space,sparse instances of data,few irrelevant features scores high over the remaining. Previous works also suggest that SVMs deliver state-of –the –art classification performance. However,success on benchmarks is a brittle justificatiom for a learning algorithm and gives only limited insight.

On the other hand,with respect to text categorisation, Artificial neural networks has fared comparable to SVMs.The performance of ANNs is statistically comparable to SVMs even with a reduced data size. ANNs have been shown as one of the best text classifiers[7]. Hence we reduced the multitude of options available for classification to ANN and SVMs.

i. Our approach

Our initial implementation used both ANNs and SVMs. But we finally used an algorithm similar to back propagation ANN .This is due to fact that SVMs are not suitable with instances where there are many well populated categories. In this system as the teacher is providing a comparatively relevant site for question extraction there is very less chance of obtaining few well populated categories.

ii. Theoretical model

In back propagation there are 2 phases in its learning cycle, one to propagate the input pattern through the network and the other to adapt the output, by changing the weights in the network. The training procedure of a back propagation is iterative, with the weights adjusted after every testing phase. The underlying principle was mainly back propagation but was slightly modified while implementing.

iii. Designed Algorithm

- Word extraction:
  Any question(sequence of characters) should be converted to words or tokens for the purpose of feature identification.This process is called as word extraction techniques depends on the language used. We have restricted to English in our system and hence word demarcation was done with the help of spaces and full stops.

- Stop word removal
  The process of removing the set of non-content bearing functional words like (what,where,is,was,etc) from the set of words produced after word extraction is called stop words removal.As these words are not unique to any particular topic, removal of these words will increase the accuracy. An exhaustive list of 400 odd stop words was used and removed from the words obtained after extraction. In text categorisation tasks it is desirable to combine morphological variants of the same words into one canonical form. However usage of stemming has its own limitations which can be elucidated with an example.Words like generous,generate,generates all reduce to gener giving absurd results.Results obtained without stemming were better comparatively in this system.

- Giving weights to terms and Training
  In the training phase the teacher goes through the questions and assigns one of the values 1 or -1 depending on whether the question belongs to the topic or not.After this assignment terms/tokens are given positive or negative weights based on whether the question that contains the word has been assigned 1 /-1 or the token is identified as essential by the teacher.The teacher is given the privilege to add words that he/she feels should necessarily belong to that particular topic and assign their corresponding weights.Provision for modification of word-weights /deletion of words by the teacher has added dynamism as well as accuracy as there is always a possibility that initial training questions may not contain all typical words of that topic.Hence a word-weight is generated at the end of the training phase which is analysed and a threshold is calculated by plotting all word-weights and the highest local minima was chosen as this will not rule out words that are important but appeared less frequently in the question data set.

- Testing phase
  Training is required only when the teacher introduces a topic to the students. Words identified as having weights above the threshold in the training phase are saved to be used in later stages.

  As testing proceeds weights of words already present in the database may get modified depending on their frequency in the subsequent sets of questions. They may also get modified if the teacher realizes that some typical words that should necessarily belong to the topic are

not present in the database. With the help of word-weight matrix weights for each question is calculated and those with weights above question threshold are classified as positive.The initial question threshold is calculated using word-weights above threshold in the training phase ,even this threshold changes with time as word-weights get modified.

III. C. Summarisation and Evaluation of Answers

Efficient summarization of answers, their automated evaluation and qualitative analysis bring us to the last part of this project. Work on summarization of large amounts of data has been effectively done and substantial results have been achieved(Interactive Multimedia Summaries of Evaluative Text ,Giuseppe Carenini , et al.,2006) .However summarization of brief answers written by students, adapted to a particular style of writing and a characteristic concept flow, requires a different approach. The idea we used behind summarization derives inspiration from the way a naive school teacher evaluates copies of her students' examination answer sheets. Just as the teacher gazes through a certain specific set of important words, the machine is trained to iterate through student answers and extract the important set of words. Relevant words from each answer are extracted and every answer is thus converted into a word matrix.

i. Related work

The algorithm we have used for summarization has derived inspiration from basic NLP techniques like Name Entity Recognition(NER), Part of Speech tagging(POS tagging) and uses concepts of ranked information retrieval, parsing and question answer techniques. However, we have modified the standard summarization algorithms to incorporate the fact that we deal with paragraphed answers that do not exceed a certain maximum limit.

Initially, the answer is fragmented and divided into multiple tokens. Then a Part Of Speech tagger is then run on each sentence and the words are tagged accordingly. However, this tagging is not accurate enough and hence we have used a novel technique for tagging based on a naturally growing resource -Wikitionary (Wiki-ly Supervised Part-of-Speech Tagging ,Shen Li et al.,2012).

Once this tagging is complete, all nouns, adjectives and verbs are put into the summarization file. If the student bolds/underlines any specific word,the word is put into the word matrix without taking its part of speech into consideration.The summarization of paragraph answers is thus complete. The case where student answers in form of bulleted points is handled slightly differently. In case the bulleted points consist of a few words, all the words are included in the summarized file along with the point number.In case the bulleted points consist of a few sentences, the sentences are summarized and referenced with respect to the point.Thus, the order of words and hence the flow of ideas is preserved and we get a lexically abridged version of student answers.

ii. Our approach

The following are all the steps that encompass the entire algorithm that runs behind our summarization. Each of the techniques described below comprise of a separate module running at the back end of our summarization module.

- **Stemming**

We use stemming to narrow our overall word matrix so as to help with the lack of similar words per answer. We have used the standard Porter's stemming algorithm for the purpose.Overall, we

see mixed results for stemming.While stemming shows improvement it also negatively affects the overall accuracy when combined with other aspects of student answers.
For eg: Operating Systems-> Operat
Operations Systems->Operat.

- **Stop Word Removal**

As in much text, that there were many common, short function words, often prepositional terms and other grammatical syntax fillers that were found specially in descriptive answers. We have not yet attempted to build a domain-specific we used a standard set of Porter stem words.

Ex: the, is, at, who, which, on
Stop words removal turned out to drastically reduce the length of passage and descriptive answers thereby giving huge accuracy boost.

- **Named entity recognition**

Entity identification is a very important part of summarization specially because it tells the system extremely relevant data about the student's answer.Though NER systems are known to be brittle, our system has been specifically designed for the student answer domain and thus it rightly extracts the necessary words depending upon the type and setting of the question.

- **Punctuation Removal**

Removing punctuation was another result of looking at our data and noticing that students have a very large variance in phrasing and word choice. This is especially true for words that may have multiple accepted forms, or words with punctuation in them. Also, because we parse the item descriptors on spaces, any punctuations that are in the phrase are left in, including ellipses, periods, exclamation points, and others. In addition, words that are concatenated are often used differently. Punctuation removal was the also an effective feature normalization method used for summarization.

- **Lowercasing**

While answering in English, students tend to have the first word capitalized. In addition, different students will capitalize different words intentionally or otherwise, depending on their intent interpretation of the word, of choice of capitalizing acronyms. This is generally not a useful normalization for the system to understand as it deteriorates the performance of the POS tagger.
Ex. President, president, CD, cd, Windows, windows

After application of the above techniques, the system generates a relevant set of words known as the summarised answer matrix for each answer. For evaluation purposes, the word matrix of student answers is compared aginst the word matrix of standard answers and depending upon the number of matches,the student is awarded marks. As the summarization process is robust and full proof,this comparison ensure maximum efficiency and the machine has to compare only the necessary and relevant parts of answers thereby reducing processing time and increasing efficiency.

IV.Implementation Details.

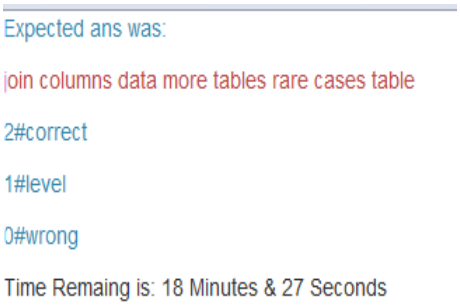
We have worked on a XAMPP PLATFORM using PHP,MySQL,JavaScript and Python as our primary languages.Web scraping has been used to scrape data (questions and answers) from the specific url. It involves the process of querying a source url, retrieving the results page and

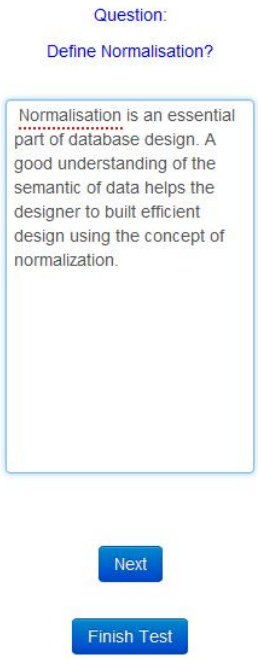
parsing the page to obtain the results. The web page to be analysed is fetched and processed in order to retrieve the data. Web data is mined using the tree-like DOM structure to analyse and describe the HTML or XML tags within the web page. Instead of defining one generic function, specific functions have been created for each website to efficiently process them.

The code for extraction has been written in PHP and JavaScript framework which adds functionality to the code. On the backend, tasks like downloading a webpage, extracting the questions from it, intelligently processing using machine learning them to get relevant questions and finally storing them in a MySQL database for future use are handled by PHP.

Though the machine learning algorithm for this was initially written in JAVA it was later adapted to PHP for increasing compatibility.Frontend tasks like displaying the questions and providing the ability to edit them in place is handled by the JavaScript Library and JQuery. Extensive care has been taken to ensure cross-browser compatibility.

There is also an option to directly add a specific question and answer pair to the database using the extension designed specifically for this purpose. If the teacher finds some interesting question while surfing the net, he can directly add it to the database by selecting the question and answer text and clicking on the extension icon. There is also an option to edit the question or answer before adding it to the database.

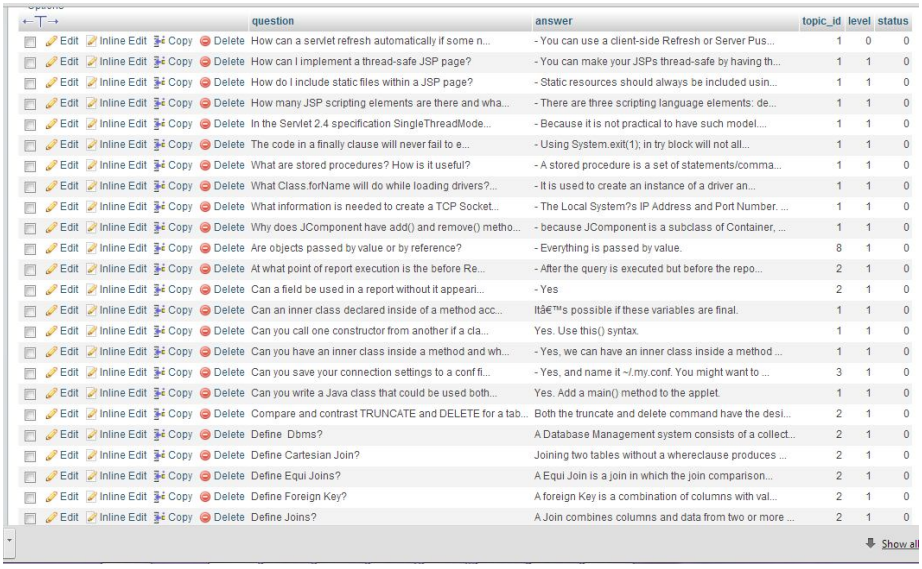




The summarization and student answer evaluation part has been implemented in Python programming language using NLTK library. Since python contain an extensive library for language processing, it stood out as the best language for development of robust modules for this project. We have developed our own POS tagger,tokeniser and hand codded features for the named entity recognition algorithm based on the dominant aspects of answers written by students.These modules when run together successfully extract the summary of an answer, evaluate it based on a standard set of answers , accordingly mark the students and thus access his performance.

II. Testing and Statistical Results

At the initial stages,as the database had questions of the order of few hundreds,there were not enough questions for testing.As the essence of machine learning lies on training the machine using maximum data ,the accuracy of classification increased as the database grew.The accuracy of classification moved from 63% to 80% as the database size moved from few hundreds of questions to thousands.Also,when the training and testing set ratio increased the accuracy further improved.When this ratio changed from 3:2 to 4:1 efficiency further improved from 81% to 89%.



	question	answer	topic_id	level	status
Edit Inline Edit Copy Delete	How can a servlet refresh automatically if some n...	- You can use a client-side Refresh or Server Pus...	1	0	0
Edit Inline Edit Copy Delete	How can I implement a thread-safe JSP page?	- You can make your JSPs thread-safe by having th...	1	1	0
Edit Inline Edit Copy Delete	How do I include static files within a JSP page?	- Static resources should always be included usin...	1	1	0
Edit Inline Edit Copy Delete	How many JSP scripting elements are there and wha...	- There are three scripting language elements: de...	1	1	0
Edit Inline Edit Copy Delete	In the Servlet 2.4 specification SingleThreadMode...	- Because it is not practical to have such model....	1	1	0
Edit Inline Edit Copy Delete	The code in a finally clause will never fail to e...	- Using System.exit(1); in try block will not all...	1	1	0
Edit Inline Edit Copy Delete	What are stored procedures? How is it useful?	- A stored procedure is a set of statements/comma...	1	1	0
Edit Inline Edit Copy Delete	What Class.forName will do while loading drivers?...	- It is used to create an instance of a driver an...	1	1	0
Edit Inline Edit Copy Delete	What information is needed to create a TCP Socket...	- The Local System?s IP Address and Port Number. ...	1	1	0
Edit Inline Edit Copy Delete	Why does JComponent have add() and remove() metho...	- because JComponent is a subclass of Container, ...	1	1	0
Edit Inline Edit Copy Delete	Are objects passed by value or by reference?	- Everything is passed by value.	8	1	0
Edit Inline Edit Copy Delete	At what point of report execution is the before Re...	- After the query is executed but before the repo...	2	1	0
Edit Inline Edit Copy Delete	Can a field be used in a report without it appeari...	- Yes	2	1	0
Edit Inline Edit Copy Delete	Can an inner class declared inside of a method acc...	Itâ€™s possible if these variables are final.	1	1	0
Edit Inline Edit Copy Delete	Can you call one constructor from another if a cla...	Yes. Use this() syntax.	1	1	0
Edit Inline Edit Copy Delete	Can you have an inner class inside a method and wh...	- Yes, we can have an inner class inside a method ...	1	1	0
Edit Inline Edit Copy Delete	Can you save your connection settings to a conf fi...	- Yes, and name it ~/.my.conf. You might want to ...	3	1	0
Edit Inline Edit Copy Delete	Can you write a Java class that could be used both...	Yes. Add a main() method to the applet.	1	1	0
Edit Inline Edit Copy Delete	Compare and contrast TRUNCATE and DELETE for a tab...	Both the truncate and delete command have the desi...	2	1	0
Edit Inline Edit Copy Delete	Define Dbms?	A Database Management system consists of a collect...	2	1	0
Edit Inline Edit Copy Delete	Define Cartesian Join?	Joining two tables without a whereclause produces ...	2	1	0
Edit Inline Edit Copy Delete	Define Equi Joins?	A Equi Join is a join in which the join comparison...	2	1	0
Edit Inline Edit Copy Delete	Define Foreign Key?	A foreign Key is a combination of columns with val...	2	1	0
Edit Inline Edit Copy Delete	Define Joins?	A Join combines columns and data from two or more ...	2	1	0

Show all

Extracting data from small document was a challenge as there was limited literature available for us. Hence, we have done extensive evaluation of the algorithm we implemented for the same. We have used the ROUGE criterion for evaluation. We started with ROUGE-1 .Though the rouge scores were impressive, this did not take care of the order of words in a sentence. Hence we resortd to a more accurate ROUGE metric-the ROUGE-L metric.

To see of there is an improvement if we increase the number of references , we also tried with 3 references for ROUGE. However, we discovered that unlike large data analysis, there was not much improvement seen in Rouge scores as the number of references increased.This observation prevails particularly for sentence level analysis.Hence, for the final version, we have used only one reference from the database. A brief table describing our results for summarisation is as follows:

	10 Answers 1 Ref	10 Answers 3 Ref	100 Answers 1 Ref	100 Answers 3 Ref	1000 Answers 1 Ref	1000 Answers 3 Ref
ROUGE-1	0.91	0.92	0.95	0.95	0.97	0.97
ROUGE-L	0.76	0.78	0.82	0.85	0.88	0.89

VII. Conclusion and Future Work:

Our approach, working with the Document Object Model (DOM) tree as opposed to regular expression matching, enables us to perform data extraction, identifying the original data and summarizing it. The techniques that we have employed, though simple, are quite effective. Currently, though we are handling many webpages, we plan to make suitable changes in our system so that any webpage can be used for data extraction in future. In future, we plan on

making our extension browser compatible and more user friendly so that it can work in any interface in which the system is installed.

Classification was currently done based on the words and weights in questions, but words that meant the same should be allocated same weights by the system and then used for classification. This can increase the credibility of the system enormously.

One main difficulty was encountered in summarisation. For answers that are extensively descriptive, students used different forms of the same word i.e. synonyms or related words. This was where the efficiency of the system declined as it failed to understand the similarity. Accessing the thesaurus corpora was possible but accessing it for each answer took vast amounts of time and was also unnecessary in most of the cases. Thus, there is trade-off between system efficiency and judicious resource management. Currently, though this issue has been handled by letting the teacher make a choice about entailing the thesaurus corpora, we plan to make modest changes in the software to handle this issue at the implementation level in future.

In conclusion,we have addressed the challenging problem of developing a fully automated intelligent testroom scenario with minimal human intervention successfully developed a system to support our claim. Generation,summarization and automated evaluation of student answers is achieved and we thus have a robust, fully functional intelligent classroom scenario in place.We feel this takes us one step closer to realising the Artificial Intelligence dream of building fully functional autonomous bots some day.

References:

[1] Irmak U and Suel T. 2006. Interactive wrapper genera-tion with minimal user effort. In Proceedings of the 15th World Wide Web. Scotland, May 23-26, 2006. PP553-563.

[2] Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S.2002. A brief survey of web data extraction tools. SIGMOD Rec. 31,2, 84

[3] Kushmerick, N.2002. Finite-state approaches to web information extraction. Proc. of 3rd Summer Convention on Information Extraction,

[4] A. F. R. Rahman, H. Alam and R. Hartono. “Content Extraction from HTML Documents”. In 1st Int. Workshop on Web Document Analysis (WDA2001), 2001.

[5] A. F. R. Rahman, H. Alam and R. Hartono. “Understanding the Flow of Content in Summarizing HTML Documents”. In Int. Workshop on Document Layout Interpretation and its Applications, DLIA01, Sep., 2001.

[6] Text Categorisation with Support Vector Machines:Learning with many relevant feactures –Thorsten Joachims

[7] Waleed Zaghloul, Sang M. Lee, Silvana Trimi, (2009) "Text classification: neural networks vs support vector machines", Industrial Management & Data Systems, Vol. 109 Iss: 5, pp.708 - 717

[8] Document Classification with Unsupervised Artificial Neural Networks-Dieter Merkl and Andreas Rauber

[9] Support Vector Machines for Text CategorizationA. Basu, C. Watters, and M. ShepherdProceedings of the 36th Hawaii International Conference on System Sciences – 2003

[10] An Extensive Empirical Study of Feature Selection Metrics for Text Classification George Forman,Intelligent Enterprise Technology Laboratory

[11] The Simple Bayesian Classifier as a Classification Algorithm -Leon Versteegen

[12] Demian Antony D'Mello,(2012), “Functional Semantics Aware Broker Based Architecture for E-Learning Web Services”, International Journal on Integrating technology in Education,, ISSN 2320-1886,Vol1,Issue1.