Pattern Matching Using Text Mining

Dr.Sujni Paul
Assistant Professor
Al Dar University College
Dubai.

Mrs.Sindhu
PhD. Research Scholar
Bharathiar University
India.

Abstract

The objective of this paper is to find the similarity between sentences using pattern based information extraction for short answer evaluation system. Pattern matching is the process of searching a sequence of tokens for the presence of the elements of a pattern. Text mining is process of extracting high quality information from text. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. This objective is employed in short answer evaluation to automate short answer grading and the patterns are recognized using pattern matching. Information extraction is carried out by text mining. The aim is to employ natural language techniques in matching sentences as well as increasing the efficiency of the algorithms.

Index terms: text mining, pattern matching, tokenization, stemming, tagging.

I. INTRODUCTION

Assessment is an important activity that is carried out in any educational process. The various modes of accessing the learner's knowledge are objective and subjective tests, quizzes assignments etc. As the objective exam usually comprises multiple choice questions, fill in the blanks and are precise to the point, subjective examsare also essential in evaluating the concepts learned by a learner. The subjective test focuses on descriptive answers which can be short answer questions or even essays. Doing the evaluation of the subjective tests automatically makes the evaluation easier. Hence developing a model for automatically grading short answers helps the examiners to evaluate the answers in less time.

II. RELATED WORK

P.Selvi,Dr.A.K.Bnerjees article"Automatic Short –Answer Grading System (ASAGS)", InterJRI Computer Science and Networking, 2010 - emphasizes on enhanced BLEU method. This method assesses a text by computing a score basedon explicit word-to-word match between the student's answerand teacher's answer. If more than one referenceis available, the matching similarity is scored against eachreference independently and the best scoring pair is used to findthe final score. Through experiments performed on a data set, we show that the semantic ASAGS outperformsmethods based on simple lexical matching; resulting up to 59 percent with respect to the traditional vector-basedsimilarity metrix.

SachinSaxena, Poonam Rani Guptainin their article, "Automatic Assessment of Short Text Answers from Computer ScienceDomain through Pattern Based Information Extraction" Proceedings of ASCNT, CDAC, 2009- system uses the information extraction whereprimarily the Part-of Speech (POS) tagger is applied on the short text answer, which istrained on Penn Treebank corpus. Then they have used metonymy technique where the nameof an object or concept is replaced with a word closely related to or suggested by theoriginal. Further patterns have been developed for each answer and classification ofpatterns is done so that marking scheme can be applied by associating marks with eachpattern to evaluate the answer in such a way that it can be found, how close it is to the right answer.

In this article "A reliable approach to automatic assessment of short answer free responses", WebLAS, Applied Linguistics & TESL, UCLALos Angeles, CA 90095 Lyle F Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J Pan, Chris Salvador, YasuyoSawaki,discusses an innovative approach to the computer assisted scoring of student responses in WebLAS (web-based language assessment system). This is a language assessment system delivered entirely over the web. Expected student responses are limited to production free response questions. The instructors and language experts do not only provide the task input and prompt, they interactively inform the system how and how much to score student responses. This interaction consists of WebLAS' natural language processing (NLP) modules searching for alternatives of the provided answer and asking for confirmation of score assignment. WebLAS processes and stores

all this information within its database, which is used in the task delivery and scoring phases.

Dr. Rama Gautam, Nikhil Bhatt, in their article "Web-based Tool for Automatic Assignment Evaluation", Special Issue of IJCCT for International Conference ACCTA-2010- examines the role of technology and Web-based software in theclassroom for evaluating a students' knowledge. The testsinvolved are subjective and the results needto be analyzed in different ways. Most approaches use eithera keyword driven approach or some form of Natural Language Processing (NLP) to analyze the results. Thispaper presents a keyword drivenapproach, backed with manual assessment.The aim is not to develop a completelyautomated system, but more like anassistant to help evaluate student knowledge. .

III. PATTERN MATCHING

Pattern matching is the problem of locating a specific pattern inside raw data. The pattern is usually a collection of strings described in some formal language. Pattern matching is used for finding the locations of a pattern within a token sequence.There are many algorithms for pattern matching. The Boyer-Moore algorithm is considered the most efficient string-matching algorithm for natural language. The BMH algorithm is a fast and easy-to-implement algorithm.

Text Mining

The text mining can be defined as the automated or partially automated processing of text. Text miningincludes pattern and knowledge discovery, deterministic and probabilistic informationextraction, trend analysis, term frequency counting, clustering and link analysisand is used to analyze unstructured information. The techniques for transforming unstructured text into structured form include tokenization, lemmatization, part-of-speech tagging and word sense disambiguation.

Short Answer Evaluation

The model to be developed is to automate short answer grading using pattern based information extraction. It is carried out through the following steps:

a. Tokenization

Tokenization is the process of breaking a text stream into words, phrases, symbols or other meaningful elements called tokens. It involves the parsing of text into separate entities, like words or punctuation marks. After the sentence is tokenized, the stop words are removed. The stop words are those that occur more frequently in a text.

b. Stemming

Stemming is the process of reducing the words to their root, stem or base form. Stemming Algorithms or stemmers have been developed to reduce a word to its stem or root form. There are many stemmer algorithms and in this paper, the porter stemming algorithm is taken since it is easy to implement.

c. Part-of-speech tagging

Part of speech tagging is the process of assigning the appropriate part of speech to a word. The common parts of speech are noun, verb, adverb, pronoun etc. The tagger returns the correct part of speech for each word in the sentence or text. There are rule based taggers and stochastic taggers. The rule based taggers requires hand written rules. Stochastic taggers use a trained corpus to assign a tag to the word. The rule based taggers are being used to perform part-of – speech tagging.

d. Word SenseDisambiguation

One word can have more than one sense. A sense is a specific meaning of word under a particular part of speech. Word sense disambiguation is the process of finding the most appropriate sense of a word in the given context. The algorithms that can be used for word sense disambiguation are dictionary-based, supervised and unsupervised algorithms. In this paper we implement dictionary based algorithm. The dictionary to be used is WordNet, the lexical online dictionary.

e. Compute similarity of sentences

To compute the similarity between sentences, we compute the similarity between synsets. A synset consists of a word, its explanation and its synonyms. The meaning of a word under one part of speech is known as a sense. The similarity between two synsets is measured using the path length similarity. The path length is obtained by counting the difference in the nodes where the synsets are located. The synsets are arranged in a tree structure in different nodes.

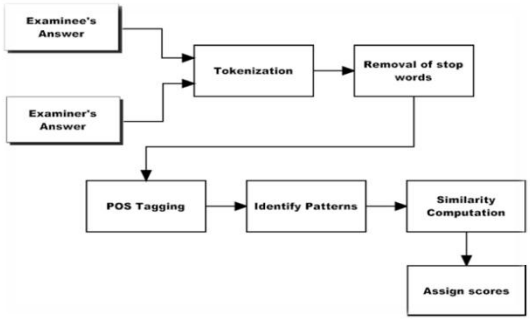

Figure1. Overall Architecture of the Short Answer Evaluation System

IV. Mathematical framework

The similarity measures used to match two patterns are:

Matching average:

This is computed by dividing the sum of similarity values of all match candidates of both sentences X and Y by the total number of set tokens.

$$\frac{2 \times Match(X,Y)}{|X| + |Y|}$$

Match(X, Y) are thematching word tokens between sentences X and Y.

Dice coefficient:

This returns the ratio of thenumber of tokens that can be matched over the total number of tokens.

$$\frac{2 \times |X \cap Y|}{|X| + |Y|}$$

V. SIMULATED RESULTS

A sentence pattern was matched with other patterns and the results obtained were plotted in a graph. The test was done to analyze the need for stemming the sentences before generating the patterns for the same. The pattern matching was performed using R. The similarity measures used were Jaccard, eJaccard, Euclidean, simple matching and cosine. The results shows that after stemming the accuracy of matches are increased. Also the similarity measure that best suits for matching is cosine.

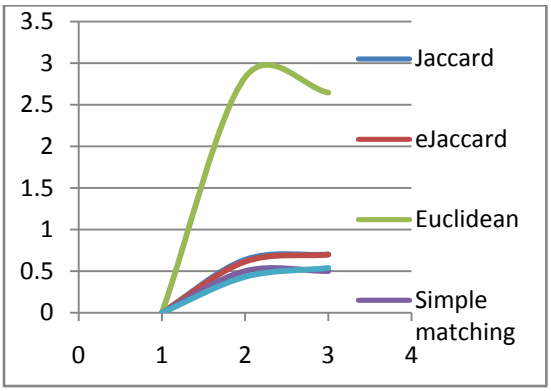

Figure 2. Before Stemming

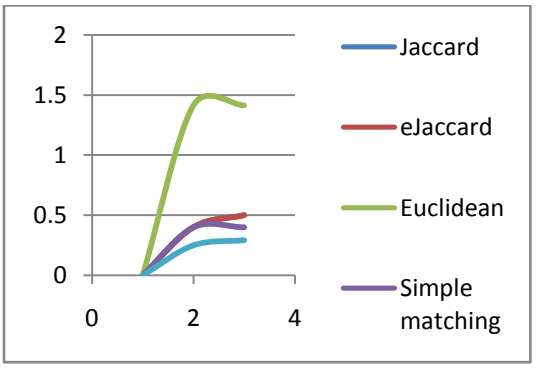

Figure 3. After stemming

VI.CONCLUSION AND FUTURE SCOPE

A model for performing pattern matching using text mining has been designed. This model can be used for short answer evaluation for more desired and accurate results. The algorithms used are porter stemming algorithm for stemming and rule based taggers for part-of-speech tagging.

The future work includes the improvisation of these algorithms for more accurate results. The efficiency of these algorithms can be increased and the disambiguation can be done in other methods.

REFERENCES

[1] P.Selvi and Dr.A.K.Bnerjee, "Automatic Short–Answer Grading System (ASAGS)" , InterJRI Computer Science and Networking, 2010.

[2] Michael Mohler and RadaMihalcea, "Text-to-text Semantic Similarity for Automatic Short Answer Grading", EACL '09 Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, 2009.

[3] SachinSaxena, Poonam Rani Gupta, "Automatic Assessment of Short Text Answers from Computer Science Domain through Pattern Based Information Extraction" Proceedings of ASCNT , CDAC, 2009.

[4] Lyle F Bachman, Nathan Carr, Greg Kamei, Mikyung Kim, Michael J Pan, Chris Salvador, YasuyoSawaki, "A reliable approach to automatic assessment of short answer free responses", WebLAS, Applied Linguistics & TESL, UCLALos Angeles, CA 90095.

[5] Dr. Rama Gautam, Nikhil Bhatt, "Web-based Tool for Automatic Assignment Evaluation", Special Issue of IJCCT for International Conference ACCTA-2010.

[6] Manu Konchady, "Text Mining Application Programming", Charles River Media, First edition, May 2006.

[7] A Text Mining Framework in Rand Its Applications, Ingo Feinerer, Department of Statistics and Mathematics, Vienna University of Economics and Business Administration, August 2008.

[8] Fast Exact String Pattern-matching Algorithms Adapted to the Characteristics of the Medical Language,ChristianLovis, MD and Robert H. Baud, PhD.

[9] Gurpreet S. Lehal, Vishal Gupta, "A Survey of Text Mining Techniques andApplications", Journal of Emerging Technologies in Web Intelligence, August 2009.

Authors Profile

Sujni Paul received MCA in P.S.G.R Krishnammal College, Coimbatore, Affiliated to Bharathiar University, India in 2000. She completed Ph.D in Computer Applications in Karunya University, Coimbatore, India in 2009. Her research interest includes Data Mining, Text Mining, E-Learning and Web Technologies. Currently she is working as Assistant Professor in AlDar University College, Dubai.

Sindhu received M.Phil in Computer Science from Bharatihiar University, Coimbatore, India in 1996**.** She is a research scholar at Bharathiar University. Her research interest includes Knowledge Sharing, Knowledge management , E-learning and Text Mining.