ATHABASCA UNIVERSITY

ASSESSING STUDENTS' ANSWERS TO OPEN QUESTIONS

BY

LAURIE CUTRONE

A Thesis project submitted in partial fulfillment

Of the requirements for the degree of

MASTER OF SCIENCE in INFORMATION SYSTEMS

Athabasca, Alberta

November, 2010

ABSTRACT

A number of Learning Management Systems (LMSs) exist on the market today. A subset of a LMS is the component in which student assessment is managed. In some forms of assessment, such as open questions, the LMS is incapable of evaluating the students' responses and therefore human intervention is necessary. This study leverages the research conducted in recent studies in the area of Natural Language Processing, Information Extraction and Information Retrieval in order to provide a fair, timely and accurate assessment of student responses to open questions based on the semantic meaning of those responses. A component-based system utilizing a Text Pre-Processing phase and a Word/Synonym Matching phase has been developed to automate the open question assessment process. A small sample of student responses were tested against the system revealing areas in which the system could be improved.

DEDICATION

This project is dedicated to my mother and my son. To my mother, Cleo, thank you for your support throughout this project. Your encouragement and genuine interest in this project has been beyond worth. To my son, Owen, it is my sincere hope that you will truly appreciate the value of education. Education opens so many doors. In my particular case, it has been life-changing. I also hope that you will see that with hard work and dedication to a task, anything can be achieved. I love you, Owen. You're my best boy.

ACKNOWLEDGEMENTS

This thesis project could not have been possible without the support of many individuals. First, I would like to thank my committee members, Dr. Jon Dron and Dr. Dunwei Wen. Your feedback, questions and sincere interest in this project are very much appreciated. Thank you. To my volunteer Human Graders, I appreciate your dedication to this project, as well as your knowledge in my testing domain. I would like to thank my family who has always been unconditionally supportive of all of my endeavours. To my good friends and 'rescuers', I could not have done this without you. My gratitude to you goes far beyond this project. Finally, I would like to thank Dr. Maiga Chang. Your encouragement and conviction in my abilities went a long way towards achieving this goal. Thank you so much for everything. I look forward to future collaborations with you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter I

The Problem

Motivation

A number of Learning Management Systems (LMSs) exist on the market today. These systems have been developed to facilitate Web-based course delivery. LMSs have improved significantly over the past decade and as a result are becoming a popular choice for not only distance education course delivery, but also for a blended style of course delivery in which a portion of the course is delivered face-to-face and a portion of the course is delivered via the Internet [3].

Learning Management Systems provide a number of advantages over face-to-face delivery. The most obvious advantage is the flexibility that the LMS provides. This flexibility can be realized by both the student and the teacher. No longer is it necessary for seats to be filled in a classroom during a specific time slot. This is especially important in situations in which the student and/or the teacher can replace their physical presence in a classroom with a virtual presence.

A subset of a LMS is the component in which student assessment is managed. In some forms of assessment – such as multiple-choice questions, the LMS software is capable of handling all of the assessment without any human intervention. In other forms of assessment – such as open questions, the LMS is incapable of evaluating the students' responses and therefore human intervention is necessary.

There are a number of commercial assessment tools on the market today; however these tools support objective-style questioning such as multiple-choice questions [14]. Multiple-choice questions will assess knowledge through the student's recall ability. However,

multiple-choice questions will not assess the learner at the higher levels of Bloom's (1956) taxonomy of educational objectives [2][6][14]. In order to assess at a higher level, it is necessary to include open-style questions in which the student is given the task as well as the freedom to arrive at a response without the comfort of recall words and/or phrases. In assessing open questions written in the traditional paper and pen format, the assessor would provide feedback directly on the student responses in order to indicate areas of correctness or incorrectness in a way of justifying the final grade given. However providing feedback using LMS software is awkward and quite time-consuming compared to the paper and pen counterpart.

Ghosh and Fatima [14] assert that with a human assessor, there is a possibility of subjectivity, time constraints, and fatigue impacting in the overall grade given to the student. Approximately 30% of a teacher's workload is spent on marking. Utilizing a system that would be capable of providing scoring within an acceptable range of that of a Human Grader, would result in a tremendous time and cost savings.

Goal and Contribution

To provide a mechanism in which the LMS software would be capable of accurately assessing the students' responses to open questions would alleviate the shortcomings described above. The system would allow for students to be evaluated at the higher levels of Bloom's (1956) taxonomy in that the students would be asked to state, suggest, describe or explain an answer [31], rather than simply recall an answer. The system would be able to provide appropriate feedback which would be absent of biases influencing the overall grade of a question. Moreover, the student responses could be evaluated in a timely manner without the need for teacher intervention.

Automating the assessment process of open questions is an area of research that has been ongoing since the 1960s [9][14][18]. However recent advances in the areas of Information Extraction [7][37][42] and Information Retrieval [8][11][17][27] have allowed for alternative approaches to be explored. Earlier work in the area of Natural Language Processing, with respect to assessing responses to open questions, focused on a statistical or probabilistic approach [20][13][18][27]. These approaches, while successful, focused heavily on *conceptual* understanding. The semantic meaning of the text was never evaluated. Rather, the location of specific words and/or phrases, and the number of occurrences of such words was being evaluated. Recent gains in Natural Language Processing have resulted in a shift in the way in which free text can be evaluated. Work in the area of Information Extraction has made significant gains in actually determining the semantic meaning of natural language text [7][37][42]. This has allowed for a more linguistic approach which focuses heavily on *factual* understanding [4][39].

This study leverages the research conducted in recent studies in the area of Natural Language Processing, Information Extraction and Information Retrieval in order to provide a fair, timely and accurate assessment of student responses to open questions based on the semantic meaning of those responses.

It should be noted that automatically assessing open questions should be done with caution. This method of assessment should not be used when the style or elegance of the student response is being evaluated [40]. Nor, at this point, should the assessment include a requirement for diagrams or examples. This form of assessment should only be used in situations in which the response is reasonably concrete, while still allowing some latitude in terms of the wording of the response.

Thesis Organization

Chapter two discusses the contributions that many researchers have made in the domains of Natural Language Processing as well as automated essay grading. Chapter three discusses the proposed methodology for this study. Chapter four focuses on the implementation and experimentation of the system. Chapter five provides an evaluation of the system following the implementation and experimentation. Chapter six provides conclusions as well as some recommendations for future work.

Chapter II

Introduction

Automatic Assessment Methods

There has been an interest in automatic assessment of open questions since the 1960s [9][14][18]. During this same era, there has been a great interest in Natural Language Processing. Natural Language Processing (NLP) involves using computers to identify semantic relations among human words [15]. It involves various dimensions of human language including grammar, usage and semantics [23]. Countless studies have attempted to decipher free text. [4] suggests work in the area of Natural Language Processing falls along a Text Technology Continuum in Natural Language Processing as shown in Figure 1 below.

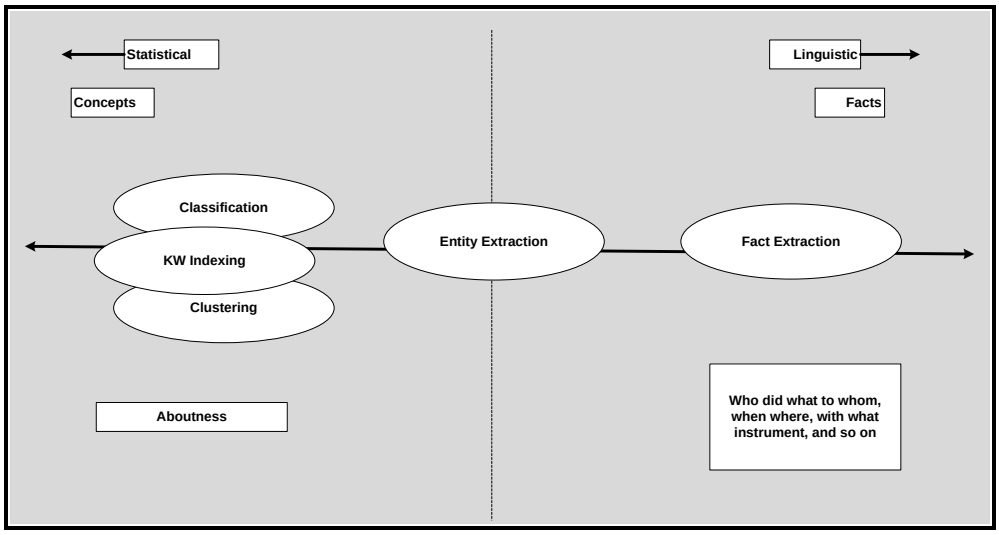

Figure 1 - Text Technology Continuum in Natural Language Processing [4]

Figure 1 illustrates that Natural Language Processing methodologies have tended to take either a statistical approach [20][13][18][27] or a linguistic approach [39]. Some studies have used a combination of approaches [5][26].

An early notable contribution to automatic essay grading was that of Project Essay Grader (PEG). [14] indicates, however, that PEG was not widely accepted because it

considered things such as the number of commas, or the number of uncommon words, yet it is notable as one of the early essay grading automated tools. PEG was an early attempt at the statistical approach to automatic essay grading.

E-Rater developed by Educational Testing Services (ETS) was a significant contribution to the assessment of open questions [14][31]. E-Rater included the structure of the text as part of the assessment process, and therefore incorporated some linguistic features. This system identified syntactic and speech features. The content of the text is compared to predefined content words. An essay that contains appropriate content words, reasonable speech features, and uses good vocabulary would receive a higher grade [31]. However, an equivalent response that failed to use the predefined content words would not receive a higher grade. This is acceptable when specific terminology must be used in the student responses. E-Rater has been successfully implemented to assess Graduate Management Admission Test (GMAT) exams since 1999 [36].

Much of the work that followed tended to continue along the statistical end of the Text Technology Continuum in Natural Language Processing (see Figure 1). Subsequent work was greatly influenced by an approach known as Latent Semantic Analysis (LSA) [13][38]. LSA uses a 'bag of words' approach in which similarity and co-location of words is evaluated [9]. LSA is a corpus-based text comparison approach and uses an algebraic technique to determine the level of similarity between the text and the corpus [13]. LSA uses word-document co-occurrences based on the corpus and presents these in a vector space [31]. LSA assumes a relationship between the meaning of text and the words used in that text. Therefore two texts that use similar words would be considered semantically similar using LSA. Texts with similar wording would be mapped closer together in the vector space [13].

This sort of approach requires a reasonable corpus to start with, and depending on the domain, the corpus may require regular updates. An additional problem inherent with LSA is that the order in which the words are presented is not considered important [9]. Therefore, the sentences: *The boy stepped on a spider.* And: *The spider stepped on a boy.* Would be considered equivalent.

Another notable technique in automatic essay grading incorporates Bayesian Networks [27]. Bayesian essay scoring uses a classification technique based on features present in the essay. This technique is supported by algebraic formulae based on calculated probabilities [27], which causes it to fall on the more statistical end of the Text Technology Continuum in Natural Language Processing (see Figure 1).

A number of approaches to automatic grading of open questions have been developed with excellent success rates (80 – 90% agreement with a human-grader gold standard) [5][10][13][18][25]. However, when leaning toward a statistical and/or probabilistic approach, there are a couple of considerations to make. First, many of these approaches require a large corpus of a previously evaluated knowledge base (such as previously graded essays) [18][38]. In many domains, the content being assessed may evolve significantly from one year to the next. This would require that the corpus be updated on a regular basis thereby potentially negating some of the time-based benefits realized from the automatic assessment tool. Secondly, using the 'closest match' and probability techniques, it is possible for a student to 'beat the system' simply by providing a number of keywords in their response, yet not accurately answering the question, or conversely, a student could answer a question accurately, yet not provide the proper keywords which results in a less-than-perfect grade.

Natural Language Processing

Despite the success rates of the automatic grading systems developed thus far, there is still an underlying problem with the past approaches. These approaches failed to attempt to equate the *meaning* of the student response to an appropriate grade. Instead, these approaches used combinations of matching algorithms, statistics, and probabilities supported by corpus and the like to make a reasonable estimate at an appropriate grade. The trend that has remained thus far in much of the previous work in automatic essay grading is that most studies have remained on the statistical end of the Text Technology Continuum in Natural Language Processing (see Figure 1). This is despite the fact that great inroads have been made in Natural Language Processing which would support an approach closer to the linguistic end of the continuum.

Information Retrieval applies a model which specifies a process in which text may be compared with specific requirements to ultimately determine the relevance of the text [19]. Information Retrieval techniques have been widely used in an effort to better understand search criteria used in general-purpose search engines [8][17]. Information Extraction involves the analysis of unrestricted text in an effort to extract relevant information. The relevant information extracted is based on some predefined guidelines [7]. Advances in both areas are of significance to this study.

Some notable IR techniques include Stemming, Chunking and the removal of Stop Words from natural language text. Stemming is an IR technique which removes suffixes in order to determine the root or stem of a word [27]. Chunking is the process of dividing sentences into noun phrases and verb groups [21]. For example the sentence: *Encryption is a mathematical formula that is applied to electronic data* would be chunked as follows: **{Encryption} is a**

{mathematical formula} that is {applied to} {electronic data}. Each *chunk* of the sentence can then be further processed based on the part of speech that the individual words represent within each chunk. Stop words are words such as pronouns, adjectives, adverbs and prepositions such as *the*, *are*, *and*, *of*, and *in* [27]. Although these words make a sentence grammatically correct, they do not contribute to the semantic meaning of the text. Studies have shown that IR has improved accuracy when stop words have been removed [27].

Information Extraction has been used in an attempt to locate text that contains a predefined semantic meaning [37][42]. Rather than using word counts, word co-location techniques, complex matching algorithms, and the like, some IE techniques have focused primarily on the actual *meaning* of the text. Many tools have emerged to assist in this effort. Generic as well as domain-specific ontologies have been developed to determine words with synonymous meaning [8][11]. Additionally, Part of Speech (POS) tagging has been incorporated to identify the various components of sentences in an effort to better understand the meaning of the sentence [12].

A significant offering to Natural Language Processing in recent years has been the development of WordNet[1] by George A. Miller of Princeton University. [30] describes WordNet as a database containing the lexical and conceptual meaning of more than 150,000 words. Words are arranged based on the relations among them. WordNet focuses on the semantic relationships between words much like a thesaurus. It allows for searching of concepts through other words that imply the same meaning. WordNet divides the words into four categories based on part of speech. These categories are nouns, verbs, adjectives and adverbs. WordNet's basic unit is the synonym set, known as the *synset*. Each synset is

[1] http://wordnet.princeton.edu/

composed of synonymous words along with pointers to related synsets. WordNet maintains both lexical and semantic relations among the synsets. Some of the relation categories include [28]:

Synonymy: Two words are considered synonyms if one can replace the other without changing the truth or value of the phrase.

Hypernymy: A hypernym is a superordinate of a word. For example, vehicle is a hypernym of car.

Hyponymy: A hyponym is a subordinate of a word and refers to an 'is a' relation. For example, a car is a vehicle.

Antonymy: Two words are considered antonyms if they are similar in all dimensions except one. For example, the words *big* and *small* both relate to size and as such are similar words, yet their meaning differs.

Meronymy: A meronym maintains a 'part of' relationship among words. For example, a door is part of a house. Door would be a meronym of house.

Holonymy: A holonym maintains a 'contains' relationship among words. For example, a house contains a door. House would be the holonym of door.

Lexical Entailment: When two verbs maintain a relationship, they are said to be lexically entailed. For example snoring entails sleeping.

Troponymy: A troponym refers to the relation among verbs that differentiates the intention or motivation of the selected verb. For example, the verbs speak and yell

have a troponymous relationship. Both convey communication, but the intention of each differs.

Part of Speech (POS) Tagging is a technique that has been widely used in Information Extraction Systems [12][16]. POS Tagging involves dividing documents into paragraphs, and then further dividing the paragraphs into sentences and phrases. Each word in each sentence is tagged with its corresponding part of speech element such as nouns, adjectives, adverbs, verbs and pronouns [12][16]. There are a number of POS Tagging tools available, some of which also perform sentence chunking to produce noun phrases and verb groups. One such tool is SharpNLP. SharpNLP, maintained by codeplex.com[2], provides a number of Natural Language Processing tools written in the C# programming language. Text is tagged using SharpNLP based on the Penn Treebank Tagset [24].

Text Pre-Processing

Much of the previous work in Natural Language Processing has incorporated a Text Pre-Processing phase in which the natural language text is prepared for the larger task of gaining a semantic understanding of the text [1][16]. Text Pre-Processing involves techniques such as chunking, stemming, removing stop words and tagging all in an effort to reduce each sentence or phrase to its canonical form.

Reducing a sentence to its canonical form is essential in order to allow for multiple, yet equivalent sentences to be considered equally when evaluating a student response to an open question. For example, the following responses would all be considered correct when responding to the question 'What is encryption?'

[2]SharpNLP http://sharpnlp.codeplex.com/ (.NET Version: http://www.codeproject.com/KB/recipes/englishparsing.aspx)

Encryption is the use of a mathematical formula that is applied to electronic data to render it illegible to anyone without a decoding key.

Encryption is the use of an algorithm applied to data making it illegible without a corresponding encryption key.

Encryption is the use of an algorithm to cause data to be illegible without a key to decode the algorithm.

Current Study

This study makes use of the recent advances in Natural Language Processing, Information Extraction and Information Retrieval to develop a system capable of automatically assessing open questions in a manner that assesses the student response based on its linguistic features. While this study focuses on linguistics, the tools utilized such as WordNet and SharpNLP do have a statistical undertone. The system reduces the supplied question, supplied answer as well as the student response to their canonical form through a comprehensive Text Pre-Processing phase. All words in the canonical form are tagged based on their part of speech. The student response and the supplied answer are then compared. In this comparison, features encapsulated within WordNet are utilized to ensure that exact word matches are not necessary in determining the level of equivalency between the student response and the supplied answer.

Chapter III

Proposed System

System Scope

The primary focus of this system is to produce software that focuses on the linguistic end of the Text Technology Continuum in Natural Language Processing (see Figure 1). In particular, the focus is on determining the semantic meaning of the student responses. This system has been developed with a goal in mind that there be a tremendous amount of flexibility in the way in which the student response is worded.

Objectives/Assumptions

1. Although grammar and spelling are critical components in learning within any domain, grammar and spelling will not be the focus of this system. Therefore, a primary assumption in this system is that all student responses as well as supplied correct answers are entered by the end users using proper spelling and grammar using complete sentences.
2. The system requires that the supplied answers as well as student responses be written in a direct manner. That is, all answers must not use analogies, slang or examples.
3. The system has been developed for the English language, and will therefore be based on English language part of speech elements.
4. For the purpose of consistency, this study focuses on a single domain of knowledge. In particular, the open questions are based on the eBusiness domain.
5. The system is supported by the generic WordNet ontology.
6. This system focuses on single-sentence responses.

7. In this early version of the software, it is assumed that each word in the reduced correct answer holds equivalent weight in the overall grade value.
8. The system was developed using a component-based architecture. See System Architecture below.

System Architecture

This system utilizes a component-based architecture. The components created in order to reduce the sentences to their canonical form are used in both the pre-processing of the supplied correct answer as well as the student response. The basic architecture of the system is shown in Figure 2 below, and is described in the following sections.

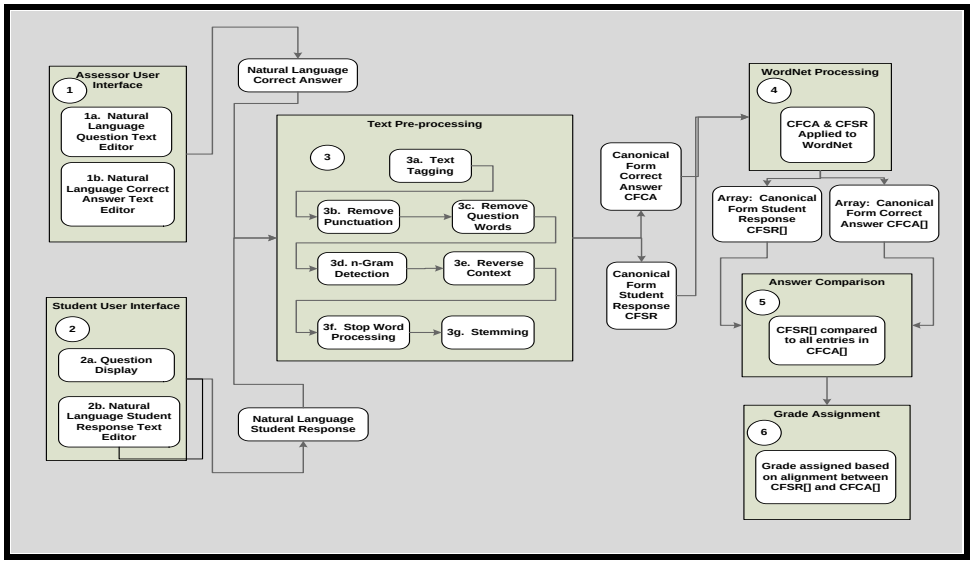

Figure 2 - Open Question Assessment Architecture

1. Assessor User Interface
    a. Natural Language Question Text Editor: Editor in which the assessor enters the open question(s) in natural language for use in student evaluation.

b. Natural Language Correct Answer Text Editor: Editor in which the assessor enters the correct answer using natural language. Note: The system assumes that the assessor will respond in complete sentences, using proper grammar and spelling (see System Scope above).

2. Student User Interface

   a. Question Display: Interface in which the student is presented with the open question(s).

   b. Natural Language Student Response Text Editor: Editor in which he student is able to use natural language to respond to the question(s) that appears in the Question Display. Note: The system assumes that the student will respond in complete sentences, using proper grammar and spelling (see System Scope above).

3. Text Pre-Processing: The Text Pre-Processing component is comprised of a number of steps which run sequentially in an effort to reduce each sentence to its canonical form [1][16]. These steps are applied to both the correct answer (CA) and the student response (SR). Additionally, a portion of these steps are applied to the Question.

In determining the canonical form of a sentence, one must apply a predefined set of rules in the pre-processing phase. In applying these predefined rules, a number of processes may take place including chunking, stemming and the removal stop words. For the purpose of assessment, an elegantly phrased sentence is reduced to its canonical form and then compared to the canonical form of the supplied answer. Therefore it must be reiterated that automatic grading of

open questions using this technique should not be used when the writing style of the text response is to be considered in the evaluation process.

a. Text Tagging: Text tagging involves applying POS tags to each word in the sentence. In addition, certain words within the sentence given an additional tag to indicate that the word is the beginning word in a 'chunk'. This will help to identify noun phrases and verb groups as necessary during later phases in the Text Pre-Processing phase. In this study, these tasks will be accomplished using SharpNLP. For example, the sentence: *Encryption is an algorithm applied to electronic data*. would be processed using SharpNLP's POS Tagger as shown in Figure 3 below:

Encryption is an algorithm applied to electronic data.
Encryption/**NN** is/**VBZ** an/**DT** algorithm/**NN** applied/**VBN** to/**TO** electronic/**JJ** data/**NNS** **./.**
NN Noun, singular **VBZ Verb, 3sg** **DT Determiner** **VBN Verb, past participle** **TO To** **JJ Adjective** **NNS Noun, plural** **. Sentence final**

Figure 3 - POS Tagging Example [22]

b. Remove Punctuation: The POS Tagger used in this project applies tags to punctuation using a different format than the tags applied to words. In order to alleviate problems associated with the punctuation tags, all punctuation is removed from the sentence.

c. Remove Question Words: In order to prevent a student from being credited for simply repeating the question words in their response, all question words are removed from the student response as well as the supplied answer. The question words to be removed are based on the canonical form of the question. As such, the question must also be subjected to a portion of the Text Pre-Processing phase prior to this step.

d. N-Gram Detection: It is important to recognize word groupings that connote a single meaning. These include compound words or proper nouns [35]. For example, the word groupings 'telephone directory' and 'National Hockey League' should not be split even though they are comprised of individual nouns that, in and of themselves, connote meaning. This pre-processing step will re-tag any identified n-grams so that the true meaning of the sentence is captured.

e. Reverse Context: Natural language text can have a variety of morpho-syntactic variations which are equivalent semantically [32]. In some cases, a sentence can be stated in a reverse form which is equivalent to a more direct approach. For example:

Encryption is a process in which a mathematical formula is applied to electronic data.

Encryption is the process of modifying electronic data by applying a mathematical formula.

Both of the sentences connote the same meaning but are phrased differently. This step looks for word combinations that indicate reverse context. In these types of sentences the reversing words are removed, and the nouns are reversed. The chunks identified in the Text Tagging step are analyzed in an effort to simplify each sentence. This provides two clear advantages to the assessment process. First each sentence is evaluated in its simplest form. Second, each sentence when being compared to the benchmark sentences takes on a similar syntactic structure which simplifies the comparison process.

f. Stop Word Processing: In this step, the tagged text is examined to determine whether any stop words exist. If so, the stop words are removed from the text [19][35]. This causes most sentences to be grammatically incorrect. However, the semantic meaning of the sentence remains. A complete sentence is shown below followed by the same sentence with stop words removed:

*Encryption **is an** algorithm applied **to** electronic data.*

Encryption algorithm applied electronic data.

g. Stemming: In the stemming phase, individual words are reduced to their canonical form or stem. The canonical form of a word is the base or lemma of that word [16][19]. For example the canonical form of the words artist and artisan is art. In order to reduce a sentence to its canonical form, the individual words within the sentence must be

examined to ensure that they are also in their canonical form. Stemming simplifies the process of locating synonyms which takes place following the pre-processing phase. For example (Note: the example is a sentence with stop words removed):

Encryption algorithm applied electronic data.

In this sentence, the canonical word for 'encryption' is 'encrypt'; the canonical form of 'applied' is 'apply'; and the canonical word for electronic is 'electron'. Therefore the sentence would be modified as follows:

Encrypt algorithm apply electron data.

4. WordNet Processing: Following Text Pre-Processing, the actual evaluation of the student responses based on the correct answer takes place. Each word in the Correct Answer in Canonical Form (CFCA) is compared to the corresponding word(s) in the Student Response in Canonical Form (CFSR). This process makes use of WordNet.NET, a .NET version of WordNet developed by Troy Simpson and maintained by Ebswift [40]. The words are first compared for an exact match. An exact match is determined based on:
    a. A matching part-of-speech tag
    b. A word match
    c. The words that have been matched have an equivalent relative position in the sentence with respect to the sentence verb(s) (if any exist).

If an exact match is found then the word in the student response is assigned a full grade value. If an exact match is not found the CFCA, as well as the CFSR are applied to WordNet. In this step, all synonyms for all words in the CFCA and CFSR are determined. A synonymous match is determined based on the same criteria as an exact match. That is, matching POS tags, synonymous words, and matching relative position in the sentences based on the sentence verb(s). If a synonymous match is found, a value is assigned to the matched word based on the level of similarity between the two matched words in the WordNet Web. The level of similarity is measured based on the relative distance between the two words within the WordNet Web. This study makes use of the WordsMatching[41] algorithm developed by Dao Ngoc Thanh in 2005 for use with WordNet.NET and maintained by The Code Project. The algorithm and supporting code traverses the WordNet synsets in an effort to determine the similarity between two words. The algorithm takes a length-based approach to determining the similarity between words. In Figure 4 below, it can be noted that the length between car and auto[motive] is 1 while the length between car and fork is 12 [41].

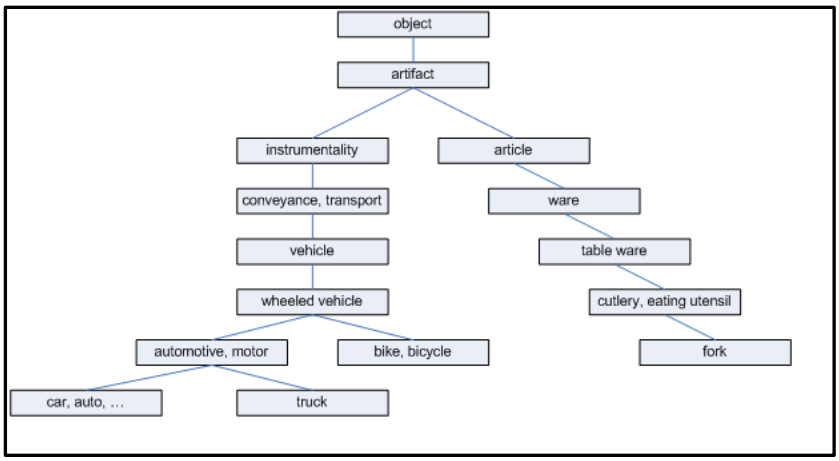

Figure 4 - Path length-based similarity measurement [41]

In general, the shorter the distance between the words, the more closely related the words. Words determined to be equivalent or synonymous according to the WordsMatching algorithm will receive a matching value of 1. Any word pairings that are not considered synonymous will receive a matching value less than 1 but no less than 0.

When applying the words ***car*** and ***auto*** from Figure 4 to the WordsMatching algorithm, a value of 1 is returned [41]. This would indicate that the words could be interchangeable in a sentence without deviating from the meaning of the sentence. As the distance between the compared words increases, the degree of synonymy decreases. This can be seen by comparing the words ***car*** and ***vehicle*** using the WordsMatching algorithm. This comparison yields a match value of 0.89 [41]. This value would indicate that the meaning of the sentence would change slightly if the words were interchanged.

When applying the words ***car*** and ***fork*** to the WordsMatching algorithm, a value of 0 is returned [41], indicating that these words are not interchangeable in a sentence, and would drastically alter the meaning of the sentence if they were interchanged.

Grade Assignment: The final grade of the open question is calculated by applying a formula that considers the number of words remaining in the canonical form of the correct answer, the cumulative WordsMatching values as well as the weight of the question. Assuming that the question has a worth of 10 marks, the student response shown in the examples above may receive a grade calculated as shown in Table 1.

Table 1 - WordsMatching Results

CACF	WordsMatching Value	SRCF
algorithm	**0.95**	formula
apply	**1.00**	employ
electron	**UNMATCHED**	process
data	**1.00**	data
Total	**2.95/4.00**	

In the results shown in Table 1, it should be noted that the words *electron* and *process* did not meet the eligibility requirements in order to be compared (See System Architecture – WordNet Processing section), and thus were not compared. Assuming the question has a worth of 10 marks, the above grade was calculated to a final result as follows:

Formula:
*Weight * (Total Response Match Values/Correct Answer Word Count)*
*10 * (2.95/4.00) = 7.375/10*

Development Methodology

A number of factors were considered when deciding on a development methodology. The development team consists of one person. As such, all analysis, design, coding, testing and implementation duties are the responsibility of a single individual. Cost for development is another consideration. Freeware and open source solutions were utilized whenever possible as long as they provide an acceptable solution. The timeline of the project is also a consideration, especially when dealing with a single developer.

A variation of the Agile development methodology was used. Agile is a 'team' based methodology, yet this project was comprised of a single member. However, the approach of 'just enough' analysis was performed to guide the developer in the various tasks. Agile is an iterative approach which supports the rapid evolution of solutions. The system is comprised of a number of independent components. These components were developed individually and as the project progressed, the components were integrated with one another when appropriate. This allowed each component to be individually developed, tested and implemented without affecting the development of the other components.

Development Tools

The development tools selected for this project were selected based on preliminary analysis of the proposed system. As the Agile development process continued, evolutions of the system emerged. During these evolutions, some development tools emerged as appropriate for the project at that particular time. Any costs associated with the development of this system were borne by the developer. The tools listed below were selected for this system. The justification of these selections is provided below.

WordNet.NET was selected as the tool used to decipher whether appropriate synonyms exist in the student response. Other WordNet-like tools exist such as the Information Content tool created by [33], as well as WordNet-like tools created for languages other than the English Language [28][29]. WordNet was selected because it is one of the largest lexical databases for the English language. Additionally, WordNet provides well-documented open source [30]. WordNet has been widely utilized in the Natural Language Processing domain, and as such, many projects were referred to while pursuing this project in an effort to best utilize the tool. As well, WordNet has a .NET version available. Finally, WordNet is a free download which satisfied budgetary concerns.

SharpNLP was selected as the POS Tagger for this project. SharpNLP was selected because it has a .NET version which will allow for integration with the .NET version of WordNet. Additionally, SharpNLP is encapsulated with a chunker which support some of the tasks required within the Text Pre-Processing phase. SharpNLP is also a free download which satisfied budgetary concerns. As well, SharpNLP is accompanied by extensive documentation as well as an active forum.

As a result of the Natural Language Processing tools mentioned above, the development tool selected for this project is C#. C# was selected based on the .NET versions of WordNet and SharpNLP. C# allows for the system to be developed as a Windows- or Web-Based application offering greater flexibility to the end-user.

In order to maintain a controlled development environment, the system was developed using a Virtual Machine. Microsoft Virtual PC 2007[3] was used to maintain this controlled

[3] http://www.microsoft.com/downloads/details.aspx?FamilyId=04D26402-3199-48A3-AFA2-2DC0B40A73B6&displaylang=en

development environment. A separate Virtual Machine environment was set up for testing the system. Microsoft Virtual PC is a free download and Microsoft.com provides extensive documentation.

Chapter IV

Experiment System and Evaluation Methods

4.1. Evaluation Methods

In order to determine the quality of the system developed, this system was evaluated using methods similar to those used in previous studies. This allowed for a better comparison of results among the various studies.

4.1.1. Gold Standard

Much of the previous work in the area of open question assessment used a benchmark standard, which is regarded as a definitive point from which to make comparisons, often referred to as the gold standard. The gold standard in previous work has been based on a comparison between the system test results and the test results utilizing one or more Human Graders [26][39].

In this study two independent Human Graders were asked to grade all student responses for all of the open test questions. The Human Graders were provided with an answer key containing a single correct answer. Both Human Graders used the single correct answer for each question as a benchmark as is common for multiple teachers to follow the same answer key when administering the same test to different groups of students. The Human Graders were provided with some guidelines in terms of the range of acceptable answers. The Human Graders logged the time spent during the evaluation process. Each Human Grader was unaware of the grading strategies and the results of the other Human Grader aside from knowing that the other Human Grader was provided with the same answer key and grading guidelines.

4.1.2. Human Grader Validation

One of the criticisms of human grading of open questions is that biases and opinions can influence the overall grade provided [14]. In an effort to prove the effectiveness of a non-biased assessment tool such as the one developed in this project, the Human Grader results were compared to determine the level of agreement among the graders.

4.1.3. System Evaluation

The system automatically assessed the same student responses as were assessed by each of the independent Human Graders. The system worked based on the same answer key as was provided to the Human Graders. The system, however, determined automatically the range of acceptable answers and appropriate deductions for answers that are at the far ends of the range. The time spent by the system was calculated in an effort to draw a comparison between the automatic assessment process versus the Human Grader (manual) assessment process.

4.1.4. Results

Results were calculated based on comparisons between the each of the Human Graders and the automatic assessment system to determine the level of agreement among the two assessment methods. Additionally, the Human Graders were compared to one another to determine the level of agreement between two humans. Results were compared based on the level of agreement, as well as the time spent among each of the grading strategies.

4.2. Prototype System

4.2.1. System Description

The experiment system has been developed as a Windows application. The system is presented in a Multi Document Interface style and depending on the logon credentials supplied, a different set of windows are supplied to the user. Three different user types ~~will~~ use the system. These types include the Assessor, the Student and Operations personnel.

4.2.2. The Assessor

The Assessor is responsible for creating the test. This includes creating and/or selecting the questions that will be included in the test. Figure 5 shows the Test Designer interface.

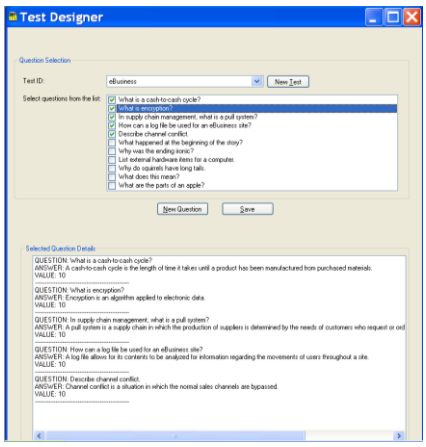

Figure 5 - Test Designer

When creating new questions, the Assessor must provide the question as well as the correct answer as shown in Figure 6. The question and the correct answer must be free of any spelling or grammar errors. In addition, the correct answer must be comprised of a single sentence.

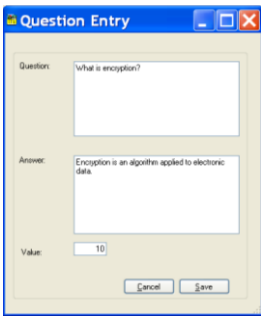

Figure 6 - Question Entry

4.2.3. The Student

The Student has the ability to take a test and review test scores. When taking a test, the student is presented with an interface that allows for navigation among the test questions. The Student is presented with the test question and an editor in which a response can be composed as shown in Figure 7. The student response must be free of spelling or grammatical errors, and must be formulated in a single sentence.

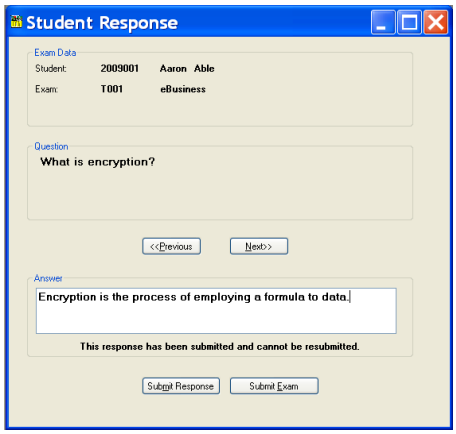

Figure 7 - Student Response Editor

When reviewing test scores, the Student is provided with the list of the questions that had appeared on the test, the student responses to those questions as well as the calculated score for each question as shown in Figure 8.

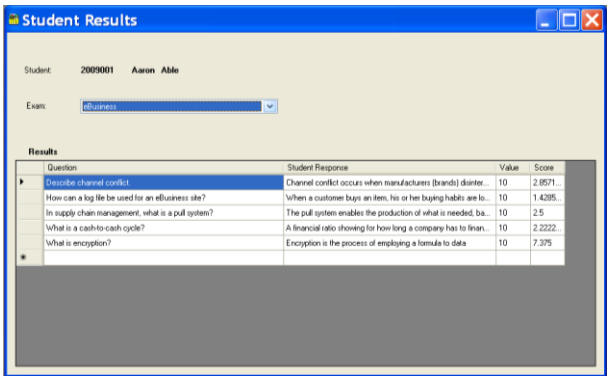

Figure 8 - Student Results

4.2.4. The Operator

In the experiment system, an Operator role was established to perform the operation of grading the tests. While this process would eventually be set up as a regular batch job, the experiment system was designed such that an end user would initiate this process. The operator selects the exam to be evaluated and manually initiates the grading process as shown in Figure 9.

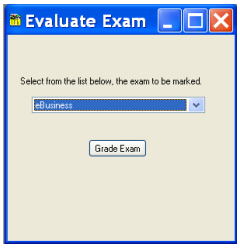

Figure 9 - Evaluate Exam

4.3. Experiment Design

4.3.1. Experiment System

The experiment system was developed under the domain of electronic commerce and mobile commerce. The open-ended questions were collected from the Internet and/or designed by the researcher. Various wikis, Google, Google Scholar, and academic databases (e.g. IEEE Xplore) were utilized to compile possible explanations and/or descriptions about

each of the questions. These explanations and/or descriptions essentially represent the student responses. Two volunteer Human Graders were provided with an answer key containing a single correct answer to each of the open questions. The Human Graders marked these explanations and descriptions found on the Internet freely and subsequently provide overall grades.

The system automatically assessed the same explanations and descriptions as were assessed by each the volunteer Human Graders.

4.3.2. Alternate Experiment System

An alternate experiment system, for further dissemination, may be developed under the domain of electronic commerce and mobile commerce. Ten volunteer students will be asked to complete a series of open questions using the interfaces provided by this system. The volunteer students will be post-secondary students over the age of 18 years. The identity of the student volunteers will not be disclosed. The students will be asked to complete their questions using a single grammatically correct sentence which is free of spelling errors. A volunteer proctor will be available to assist in the formulation of the sentences to ensure that the spelling and grammar guidelines are met. The proctor will not be permitted to provide suggestions in terms of the semantic correctness of the student responses. The questions that the students will be asked can be seen in Appendix A.

Two volunteer Human Graders will be asked to grade each of the student responses. The Human Graders will be educational professionals with experience in grading open questions as well as knowledge in the assessed domain. Each Human Grader will be provided with an answer key containing a single correct answer to each of the open questions. The answer keys provided to the Human Graders will be identical. The answer

key will be accompanied by grading guidelines for each of the Human Graders to follow. The Human Graders will be given a certain degree of latitude when determining the correctness of the student responses, and subsequently the overall grade provided to the student. The answer key can be seen in Appendix B. The grading guidelines can be seen in Appendix C. The Human Grader time log sheet can be seen in Appendix D.

4.3.3. Hypotheses

When evaluating the system, it was measured based on performance, consistency and accuracy. The Human Graders were also assessed in the same categories. As well the internal functionality of the Auto-Assessor system was assessed in order to evaluate the architecture and internal algorithms. The following list describes the outcome expectations of the experiment as well as the expectations of the performance within the inner functionality of the Auto-Assessor System.

4.3.3.1. Speed

It was expected that the Auto-Assessor System would complete the assessment process measurably faster than both Human Graders.

4.3.3.2. Consistency

It was expected that the Auto-Assessor System would show consistent marking among all student responses.

It was expected that the Human Graders would show a measurable degree of inconsistency between each other.

It was expected that individual Human Graders would show a measurable degree of inconsistency among multiple responses to the same question.

It was expected that the Auto-Assessor System would maintain a level of agreement which was equivalent to that of the Human Graders with the same or better level of agreement as that shown between the two Human Graders.

4.3.3.3. Accuracy

It was expected that the Auto-Assessor System would accurately grade each response based on the key provided.

It was expected that the Human Graders would accurately grade each response based on the key provided.

4.3.3.4. Text Pre-Processing

It was expected that the Auto-Assessor System would correctly reduce each sentence to its canonical form. That is, the sentences would be reduced to only those words that provided the semantic meaning of the sentence.

4.3.3.5. WordNet Processing

It was expected that the WordNet database would contain a reasonable number of synonymous word options thereby allowing the end user (student or teacher) latitude when composing their response.

Chapter V

Evaluation and Discussion

Evaluation

In evaluating the system, data was gathered which would allow for measurement in the areas of speed, consistency, accuracy, Text Pre-Processing performance, WordNet processing performance and WordsMatching performance. Table 2 shows the grades produced by each of the Human Graders as well as the Auto-Assessor System for each of the five questions in each of the five tests. Figure 10 and Figure 11 show a graphical representation of the data. (Note: The questions, correct answers and student responses are shown in Appendixes A, B and F respectively).

Table 2 - Grade Distribution by Question

Grade Distribution by Question				
		Human Grader 1	Human Grader 2	Auto-Assessor
Question 1	Student A	5	7	2
	Student B	10	7	2
	Student C	6	6	3.25
	Student D	5	5	2
	Student E	10	10	5.75
Question 2	Student A	8	8	1.5
	Student B	10	10	5.5
	Student C	5	4	0
	Student D	7	7	0
	Student E	8	8	3.5
Question 3	Student A	10	5	2.833333333
	Student B	10	6	4.5
	Student C	9	8	5
	Student D	7	4	2.833333333
	Student E	8	6	2.833333333
Question 4	Student A	6	8	0.928571429
	Student B	6	5	0.928571429
	Student C	5	4	0.928571429
	Student D	0	3	0.928571429
	Student E	5	2	0
Question 5	Student A	10	10	1.166666667
	Student B	0	0	0
	Student C	10	10	2.833333333
	Student D	4	4	4.5
	Student E	0	1	2.833333333

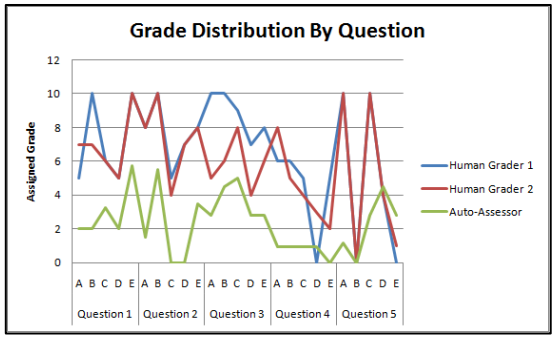

Figure 10 - Grade Distribution by Question

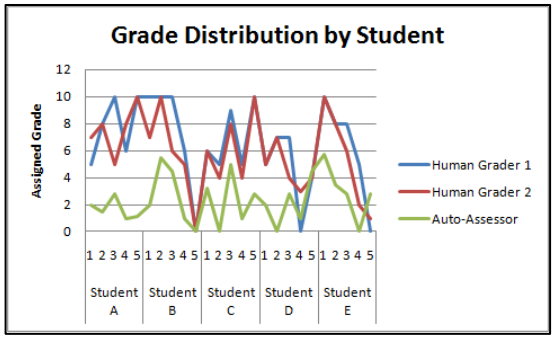

Figure 11 - Grade Distribution by Student

Figure 10 and Figure 11 reveal two significant trends within the data. First, it is clear that the Auto-Assessor system consistently provided a lower grade than both of the Human Grader counterparts. Second, it can be seen from the Figures that despite the lower grades given by the Auto-Assessor System, there seems to be a degree of correlation between the Auto-Assessor System and Human Grader 1. That is, when Human Grader 1 provided a higher grade, so did the Auto-Assessor System. This can be more clearly seen when removing Human Grader 2 from the chart as shown in Figure 12 below. This suggests that by refining the Auto-Assessor System, a better agreement rate could be achieved.

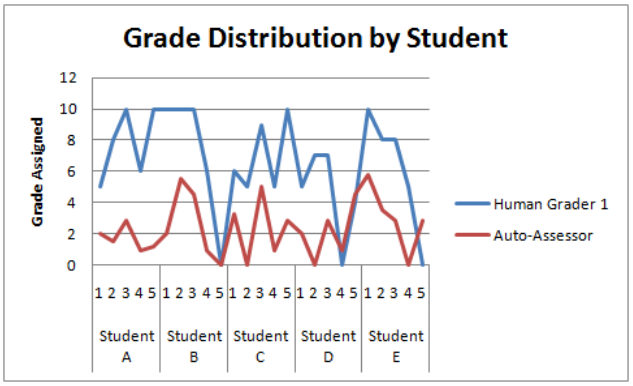

Figure 12 - Auto-Assessor and Human Grader 1 Comparison

5.1. Speed (Hypothesis 4.3.3.1)

Table 3 shows the time spent grading each of the tests by each of the Human Graders as well as the by the Auto-Assessor System. This data is also represented in Figure 13 below. The data clearly shows a significant time gain when using the Auto-Assessor System. When fully implemented this time-based improvement could result in a significant decrease in workload hours allocated to assessment.

Table 3 - Time Spent Grading (in minutes)

Time Spent	Student A	Student B	Student C	Student D	Student E	Total
Human Grader 1	5.00	6.00	2.50	5.00	4.00	22.50
Human Grader 2	3.25	2.83	1.95	2.83	2.83	13.69
Auto-Assessor						1.37

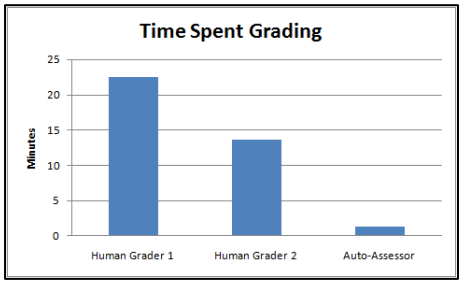

Figure 13 - Time Spent Grading

5.2. Consistency (Hypothesis 4.3.3.2)

The Human Graders exhibited a reasonable level of agreement. As shown in Table 4, the Human Graders arrived at scores within 10% (1 mark) of each other in 16/25 (64%) responses, with all scores for Student C falling within 10% of one another (see Table 2).

Table 4 - Human Grader Agreement Rate

	Human Grader Agreement Rate			
	Same Score	Within 10%	Within 20%	Greater than 20%
Number of Responses	11	5	3	6

As shown in Table 5, the Auto-Assessor System consistently (86% of graded responses) arrived at a grade lower than Human Grader scores.

Table 5 - Auto-Assessor vs. Human Grader Agreement Rate

	Auto-Assessor vs. Human Grader Agreement Rate	
	Auto-Assessor Produced Lower Grade	Auto-Assessor Produced Equal or Higher Grade
Number of Human Graded Responses	43	7

In 3 of 25 (12%) responses, see Table 6, the Auto-Assessor System scored the test within 10% (1 mark) of one or both of the Human Grader scores.

Table 6 - Auto-Assessor vs. Human Grader Agreement Rate Percentages

	Auto-Assessor vs. Human Agreement Rate			
	Same Score	Within 10%	Within 20%	Greater than 20%
Number of Responses	1	2	3	19

5.3. Accuracy (Hypothesis 4.3.3.3)

Question 4 allowed the greatest room for interpretation. The question was "How can a log file be used for an eBusiness site?" Each (simulated) student gave an *example* of situations in which a log file may be used; however the question required a more general answer. That is, the question did not ask the student to "Give an example of how a log file can be used for an

eBusiness site." In grading this question, the Auto-Assessor System consistently graded each response, providing a grade below 1/10 in all cases. For this question, the Human Graders allowed this deviation from the key, despite being given instructions to follow the answer key. In one case, Human Grader 2 awarded a student (Student A) a grade of 8/10 despite the fact that the response gave a specific situation in which a log file may be used (i.e. "when a customer buys an item"). The same Human Grader disallowed an example given by Student E providing a grade of 2/10. This shows a marked degree of inconsistency among grades provided by the same Human Grader. As well, Question 4 showed a great deal of disparity among the two Human Graders; in two cases this disparity was 30%. Figure 14 shows a graphical representation of Question 4. In this sort of situation, that is, a situation in which the Human Grader deviates from the key, the Auto-Assessor System consistently disallowed examples thereby exhibiting a consistent grading strategy to all responses.

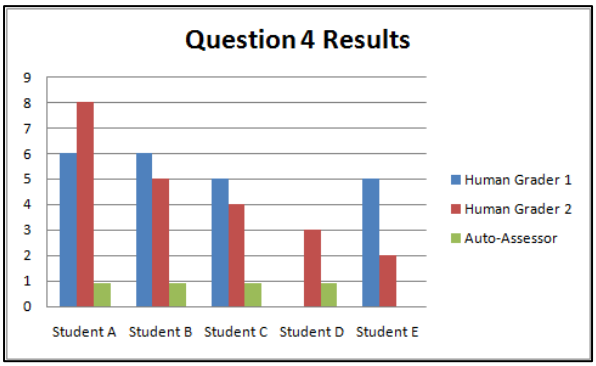

Figure 14 - Question 4 Results

5.4. Text Pre-Processing (Hypothesis **Error! Reference source not found.**)

When taking a closer look at the data, the canonical forms of the sentences were examined. The canonical form was examined to ensure that only those words that contributed to the meaning of the sentence remained. The results of this examination are

shown in Appendix F. As the data clearly shows, the system did not always remove all of the insignificant words in the sentences. Figure 15 shows the net result of including the insignificant words in the overall evaluation of the student response. Note in the figure that the correct answer should have been reduced to two words but in actuality was reduced to six words. This resulted in four additional (insignificant) words requiring matching. In this particular case the four additional words were not matched resulting in a decrease in the value of the student response by 4 marks or 66.7%.

5.5. WordNet Processing (Hypothesis 4.3.3.5)

In addition, when looking at the canonical forms of the sentences, WordNet was queried to determine whether the remaining words in each of the sentences were in fact represented in WordNet. Looking again at Figure 15 it can be noted that the word disintermediate was not found in the WordNet database. In fact, disintermediate is not found in the standard dictionary provided by Microsoft Office. This is a word specific to the eBusiness domain. Since this word was not found in the WordNet database, the match was not detected by the WordsMatching algorithm resulting in a decrease in the overall grade by 16.7%. Further analysis of the WordNet processing revealed that six domain-specific words were not represented in the WordNet database, or were represented outside the context of the eBusiness domain.

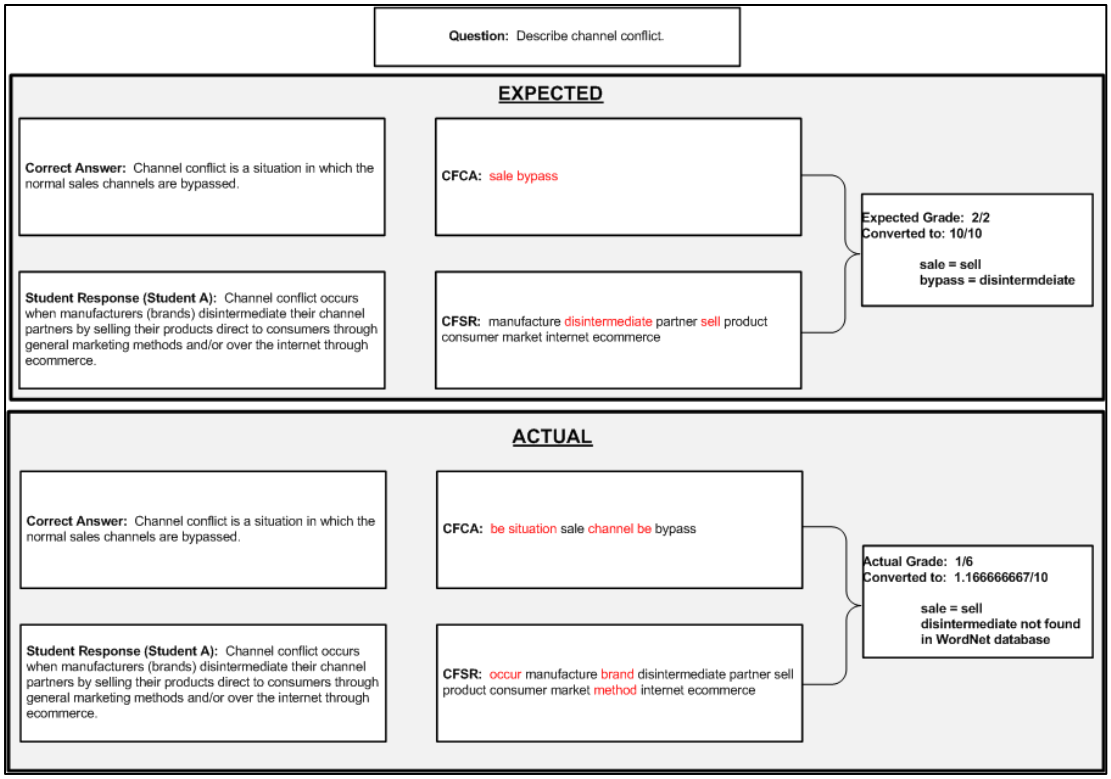

Figure 15 - Canonical Form Evaluation

Discussion

Noteworthy findings were revealed when evaluating the data. The analysis of the data has shown that despite a reasonable agreement level among the Human Graders, it should be noted that the Human Graders disagreed by more than 10% in 9/25 (36%) of the graded responses. This clearly shows a need for a more consistent grading process. Additionally, the Human Graders took some liberties when grading the student responses. This was especially true when grading Question 4. This is of significance when student responses are implied or inferred. Some Human Graders may recognize the inference, while others may not. This could lead to inconsistent grading.

When evaluating the Auto-Assessor System, it was revealed that there are two significant areas which require improvement. First, the stop words processing component of the Text

Pre-Processing phase requires additional work to ensure that all stop words are removed. Second, in some cases not all domain-specific words exist in the WordNet database. This implies the need to augment the WordNet database with domain-specific terms.

Comparison to Previous Work

With the volume of previous work available [5][13][18][20][27][39], it is important to determine how the Auto-Assessor System compares to its predecessors in the area of automatic essay grading. Noted comparisons are listed below.

The sample (25 responses) in this study was extremely small in comparison to other studies with 638 responses in [5] and 448 responses in [26]. A larger sample may have provided for more definitive results.

The Auto-Assessor system graded the 25 student responses in 1.37 minutes with a mean time spent being 3.82 seconds per response. In [13], the students received feedback within 20 seconds, however the length of the response must be considered here. In the Auto-Assessor System test, (simulated) students were limited to a single sentence response, while in [13], students were free to provide a lengthier response. In both cases, the need for a server to process responses is evident.

The Auto-Assessor System supports previous work in that it recognizes the value of a Text Pre-Processing phase [6][14][20][39], as well as the use of WordNet [39] within the system architecture.

Much of the previous work was validated against a Human Grader gold standard [5][14][26]. In these studies, the agreement rate between the systems and the Human Grader(s) ranged between 89%[5] and 92.5% [26]. The agreement rate of the Auto-Assessor

System failed to reach such levels of agreement. The Auto-Assessor System agreed with the Human Graders in only 7 of the collective 50 graded responses or a 14% agreement rate.

Chapter VI

Summary and Future Work

Summary

Auto-Assessor is a system that leverages Natural Language Processing tools including WordNet.NET and SharpNLP in order to evaluate student responses to open questions. This system differs from much of the previous work in open question assessment in that it focuses on the linguistic end of the Text Technology Continuum in Natural Language Processing (see Figure 1). The goal of the system was to produce accurate, consistent grades for student responses to open questions by deciphering the semantic meaning of the response. In addition, the system allows for a great deal of latitude when composing the responses rather than requiring specific key words.

The development of the system used a variation of an Agile methodology. This methodology complemented the component-based architecture. Each component was initially created using the 'just enough' standard supported by the Agile methodology. This allowed for frequent incremental tests of the system in which problems could be identified early. The component-based architecture selected seems appropriate as the ultimate goal of the system would be to support existing Learning Management Systems. As such, only those components required could be 'plugged into' an existing LMS to allow for open-ended question assessment.

When evaluating the system, it became quite clear that the domain in which the system was developed contained a number of words that are not supported by the WordNet database. As such, mis-grading did occur. Had a more 'generic' domain been selected for evaluation purposes, the results would have been somewhat more convincing. Conversely, it is expected

that the results of the system evaluation could be improved by coupling WordNet with a domain-specific ontology in an effort to ensure that all domain-specific words are effectively evaluated.

Previous work in this area has delivered impressive results, however these results must be improved upon in order to gain user acceptance. While the Auto-Assessor System test resulted in less-than-satisfying results, it does contain components that could be integrated with previous work. For example, the reverse-context component within the Text Pre-Processing phase (see System Architecture) could be used to alleviate the problem identified with systems the followed the LSA approach [13][38]. As well, the WordNet processing phase (see System Architecture) could be used to augment systems such as E-Rater in which exact word matches are required [31]. As well, this system may be used to support existing systems by providing formative assessment feedback to the teacher such that instruction may be adjusted to meet the learner's needs, such as was the case with [39].

Future Work

The approach used to develop this system is one that leans toward the linguistic end of the Text Technology Continuum in Natural Language Processing (see Figure 1). This is a relatively novel approach that is new to the Automatic Essay Grading domain. As such, there are many areas of which this system can be improved. These improvement areas are divided into two categories. The first category identifies critical items that must be addressed before the system can move forward with the second category which identifies items that would make the system more practical for mainstream acceptance. The critical items, once addressed will provide for a more accurate system. At that point the practical items should be addressed.

Critical Items

1. The top priority for future work with this system lies in improving the Text Pre-Processing phase. The Text Pre-Processing phase of the system requires additional work to ensure that only those words that contribute to the semantic meaning of the sentence remain. Specifically the stop word processing component requires additional algorithms, or additional passes through the existing algorithms in order to ensure that all insignificant words are removed. Improvements to the Text Pre-Processing phase will result in grading that more closely aligns with the Human Grader counterparts. This will go a long way in gaining user acceptance to this system.
2. It is important that the WordNet Processing phase be modified such that all categories of synonyms be identified and presented to the WordsMatching algorithm [41].
3. It is also recommended that the system undergo more extensive testing. In this study, five separate answers to each question were used to perform the evaluation. Given the multitude of variations of equivalent sentences in the English language, the system cannot be effectively evaluated using only five responses to each question. In the future, it would be recommended that a much larger pool of responses be used in order to perform a more accurate and extensive evaluation. The larger pool of responses may also uncover further areas in which improvements to the system could be made.

Practical Items

1. In future systems, it is recommended that a spell checker and grammar checker be incorporated. These checking utilities could be incorporated to assist the student in formulating their response. Conversely, these utilities could be incorporated in the assessment process in which a deduction to the final grade would be included if the response was not free of spelling or grammatical errors. Ultimately, it is desired that the assessor have the freedom to decide on a test-by-test basis whether the spelling and grammar checking be provided to the student or used as an assessment requirement. By incorporating a spell checker and grammar checker, the evaluation process will have an assurance that the questions have been formulated using proper spelling and grammar, allowing the current assumption to be removed.
2. As the system emerges, it will become necessary to augment the synsets provided by WordNet with domain-specific knowledge. Previous work has used ontologies to support the additional words that are domain-specific. It is recommend that ontologies be incorporated to the system as domain-specific needs emerge. The inclusion of domain-specific words will allow this system to be utilized within any domain. In developing domain-specific ontologies, Natural Language techniques such as semantic interpretation could be utilized to ensure appropriate equivalencies between words within a specific domain can be identified. As well, named-entity recognition may be used to identify domain-specific terms within a document. In doing so, a word could be appropriately mapped to either the

generic WordNet database or the more domain-specific ontology for synonym lookup.

3. The system will be of great value when it is not limited to a single sentence response. It is recommended that in future versions, the length of the response would have no limit, except any limit expressed by the assessor. Future versions of this software should allow for multi-sentence and multi-paragraph responses where the collective meaning of all sentences could be evaluated. The appeal of an open question is that it allows the user to freely come up with a response. By evaluating multiple sentences for their collective meaning, a great deal more flexibility will be provided to the users of the system.
4. In cases in which a lengthy answer is required, future versions of this system should allow the assessor to apply varying weights to the various portions of the student response. That way each sentence is not given the same emphasis and certain sentences and phrases would stand out as key components to the ultimate grade given. Allowing varying weight values ultimately will allow for a more accurate grade assigned.
5. An addition that would be desired for future versions of this software is the ability for the assessor to flag certain keywords that are required for the response. These keywords would be flagged in such a way that alternate synonyms not be generated. The ability to flag the key words would be worked into the Assessor user interface. This additional feature would be of great value in domains in which the proper use of terminology is essential.

6. Future work is recommended to eventually integrate this assessment tool with a Learning Management System (LMS). This would allow the students to be evaluated using an interface that they are familiar with. Integrating with the LMS will involve developing the system as a plug-in. This would involve the Text Pre-Processing and the WordNet Processing components to reside on the same server as the LMS. As well, a component would be required so as to interface with the host LMS.
7. Future work may also reveal an opportunity to the integrate additional Natural Language Processing techniques in order to more accurately decipher the semantic meaning of the sentence.

References

[1]. Abdul Seoud, R. A., Youssef, A-B.M., and Kadah, Y. M. (2007). Extraction of Protein Interaction Information from Unstructured Text using a Link Grammar Parser. *Proceedings of the International Conference on Computer Engineering & Systems, ICCES'07*, 70-75.

[2]. Bacon, D. R. (2003). Assessing Learning Outcomes: A Comparison of Multiple-Choice and Short-Answer Questions in a Marketing Context. *Journal of Marketing Education*, 25(1), 31 – 36.

[3]. Baggaley, J. (2008). Where did Distance Education Go Wrong? *Distance Education*. 29(1), 39 – 51.

[4]. Bean, D. (2007). How Advances in Search Combine Databases, Sentence, Diagramming and 'Just the Facts'. *IT Professional*, 9(1), 14-19.

[5]. Burstein, J., Kukich, K., Wolf, S., Lu, C., Chodorow, M., Bradenharder, L., and Harris, M.D. (1998). Automated Scoring using a Hybrid Feature Identification Technique. *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 206-210.

[6]. Callear, D., Jerrams-Smith, J., and Soh, D. (2001). CAA of Short Non-MCQ Answers. *Proceedings of the 5th International Computer Assisted Assessment Conference (CAA01),* Loughborough.

[7]. Chang, H-H., Ko, Y-H., and Hsu, J-P. (2000). An Event-Driven and Ontology-Based Approach for the Delivery and Information Extraction of E-mails. *Proceedings of the International Conference on Multimedia Software Engineering 2000,* 103-109.

[8]. Chien, B-C., Hu, C-H., and Ju, M-Y. (2007). Intelligent Information Retrieval Applying Automatic Constructed Fuzzy Ontology. *Proceedings of the International Conference on Machine Learning and Cybernetics*, 4(19-22), 2239-2244.

[9]. Datar, A., Doddapaneni, N., Khanna, S., Kodali, V., and Yadav, A. (2004). EGAL – Essay Grading and Analysis Logic, unpublished.

[10]. Dessus, P., Lemaire, B., and Vernier, A. (2000). Free-Text Assessment in a Virtual Campus. *Proceedings of the 3rd International Conference on Human-Learning Systems*, 61-75.

[11]. Dridi, O. (2008). Ontology-Based Information Retrieval: Overview and New Proposition. *Proceedings of the 2nd International Conference on Research Challenges in Information Science 2008, RCIS 2008*, 421-426.

[12]. Dung, T. Q., and Kameyama, W. (2007). A Proposal of Ontology-based Health Care Information Extraction System: VnHIES. *Proceedings of the 2007 International Conference on Research, Innovation and Vision for the Future*, 1-7.

[13]. Foltz, P., Laham, D., and Landauer, T. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer Enhanced Learning*, 1(2).

[14]. Ghosh, S., and Fatima, S.S. (2008). Design of an Automatic Essay Grading (AEG) System in Indian Context. *Proceedings of the IEEE Region 10 Conference, TENCON 2008*, 1-6.

[15]. Girju, R., & Badulescut, A., and Moldovan, D. (2006). Automatic Discovery of Part-whole Relations. *Computational Linguistics*. 32(1). 83 – 135.

[16]. Hwang, M., Baek, S., Choi, J., Park, J., and Kim, P. (2008). Grasping Related Words of Unknown Word for Automatic Extension of Lexical Dictionary. *Proceedings of the 1st International Workshop on Knowledge Discovery and Data Mining*, 31-35.

[17]. Kang, K., Lin, K., Zhou, C., and Guo, F. (2007). Domain-Specific Information Retrieval based on Improved Language Model. *Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, 2, 374-378.

[18]. Larkey, L. S. (1998). Automatic Essay Grading using Text Categorization Techniques. *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*. 90-95.

[19]. Lee, J.-W. (2007). A Model for Information Retrieval Agent System. *Proceedings of the 2007 International Conference on Multimedia and Ubiquitous Engineering (MUE '07)*, Seoul, South Korea. 413 – 418.

[20]. Li, B., Lu, J., Yao, J.-M., and Zhu, Q.-M. (2008). Automated Essay Scoring using the KNN Algorithm. *Proceedings of the 2008 International Conference on Computer Science and Software Engineering*, 1(12-14), 735-738.

[21]. Liao, P., Liu, Y., and Chen, L. (2006). Hybrid Chinese Text Chunking. *Proceedings of the 2006 IEEE International Conference on Information Reuse and Integration*. 561-566.

[22]. LingPipe Part-of-Speech Demo. LingPile, available online at http://lingpipe-demos.com:8080/lingpipe-demos/pos_en_general_brown/textInput.html.

[23]. Link Grammar Parser. Computer Software. Abiword, available online at http://www.abisource.com/projects/link-grammar/.

[24]. Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313-330.

[25]. Ming, Y., Mikhailov, A., and Kuan, T. (2000). Intelligent Essay Marking System. *Learners Together*, NGEE ANN Polytechnic, Singapore.

[26]. Mitchell, T.; Russell, T.; Broomhead, P. and Aldridge, N. (2002). Towards Robust Computerised Marking of Free-Text Responses. *Proceedings of the 6th Computer Assisted Assessment Conference*, Loughborough.

[27]. Rudner, L., and Liang, T. (2002). Automated Essay Scoring Using Bayes' Theorem. *Journal of Technology, Learning, and Assessment*, 1(2).

[28]. Sahoo, K., and Vidyasagar, V. E. (2003). Kannada WordNet – A Lexical Database. *Proceedings of TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*, 4, 1352-1356.

[29]. Shim, K-S., Ock, C-Y., Kim, D-M., Choe, H-S., and Kim, C-H. (2008). Finding Similar Texts using U-Win. *Proceedings of the International Conference on Advanced Language Processing and Web Information Technology*, 43-48.

[30]. Sosa, E., Lozano-Tello, A., and Prieto, A. E. (2008). Semantic Comparison of Ontologies based on WordNet. *Proceedings of the 2008 International Conference on Complex Intelligent and Software Intensive Systems CISIS 2008*, 899-904.

[31]. Sukkarieh, J. Z., Pulman, S. G., and Raikes, N. (2003). Auto-marking: Using Computational Linguistics to Score Short, Free Text Responses. *Proceedings of the 29th International Association for Educational Assessment (IAEA)*, Manchester.

[32]. Szpector, I., and Dagan, I. (2007). Learning Canonical Forms of Entailment Rules. *Proceedings of the International Conference on Recent Advantages in Natural Language Processing (RANLP)*, Bulgaria.

[33]. Temperley, D., Sleator, D. D. K., and Lafferty, J. (2008). Link Grammar, available online at http://www.link.cs.cmu.edu/link/index.html.

[34]. Trites, G., Borit, J. E., and Pugsley, D. (2006). E-Business A Canadian Perspective for a Networked World Second Edition. Pearson Education Canada Inc., Toronto, Ontario.

[35]. Vallez, M., and Pedraza-Jimenez, R. (2007). Natural Language Processing in Textual Information Retrieval and Related Topics. Available Online. www.hipertext.net, 5(2007).

[36]. Wang, J., and Stallone, B. M. (2008). Automated Essay Scoring versus Human Scoring: A Correlational Study. *Contemporary Issues in Technology and Teacher Education*, 8(4). 310-325.

[37]. Wang, H., Yuan, L., and Shao, H. (2008). Text Information Extraction Based on OWL Ontologies. *Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery FSKD*, 4, 217-222.

[38]. Wiemer-Hastings, P., Allbritton, D., and Arnott, E. (2004). RMT: A Dialog-Based Research Methods Tutor with or without a Head. *Proceedings of the 7th International Conference Intelligent Tutoring Systems, LNCS 3220*, Springer, 614-623.

[39]. Williams, R., and Dreher, H. (2004). Automatically Grading Essays with MarkIT. *Proceedings of Informing Science Conference*, Australia, 25-28.

[40]. WordNet.NET. Computer Software. Ebswift, available online at http://opensource.ebswift.com/WordNet.Net/.

[41]. WordsMatching. Computer Software. The Code Project, available online at http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx?msg=1755999.

[42]. Zhu, Q., and Cheng, X. (2008). The Opportunities and Challenges of Information Extraction. *2008 International Symposium on Intelligent Information Technology Application Workshops* (IITAW 2008), 597-600.

Appendix A

Assessment Questions

1. What is a cash-to-cash cycle?
2. What is encryption?
3. In supply chain management, what is a pull system?
4. How can the use of a log file assist in attracting new customers to an eBusiness site?
5. Describe channel conflict.

Appendix B

Assessment Key

1. What is a cash-to-cash cycle?

A cash-to-cash cycle is "the length of time from purchasing materials until a product is manufactured."[34]

2. What is encryption?

Encryption is "the use of a mathematical formula that is applied to electronic data to render it illegible to anyone without a decoding key." [34]

3. In supply chain management, what is a pull system?

A pull system is "a supply chain in which the production of suppliers is determined by the needs of customers who request or order goods, necessitating production."[34]

4. How can the use of a log file assist in attracting new customers to an eBusiness site?

A log file allows for its contents to be analyzed for "information regarding the movements of users throughout a site". [34]

5. Describe channel conflict.

Channel conflict a "situation in which various sales channels for a single organization operate in competition with each other". [34]

Appendix C

Human Grader Guidelines

- Answer must be formulated into a single sentence. If the answer spans more than one sentence, do not provide a grade.
- Assessor may accept any answer or portion of an answer that is synonymous with the key provided.
- Answers that are not quite synonymous but connote the same general meaning are acceptable, but should include a reasonable deduction for the lack of accuracy.
- Deductions can range from the full weighted score of an answer to portions of marks comprising the full-weighted score, that is, part marks.
- Each question must be given a grade within a range including a maximum of 10 marks and a minimum of 0 marks.
- Human Graders are not to discuss grading results with other Human Graders taking part in this study.

Appendix D

Verification of Questions and Answer Key

Name: Click here to enter text.

Title: Click here to enter text.

Organization: Click here to enter text.

I certify that the answer key listed in Appendix B of the Thesis paper ASSESSING STUDENTS' ANSWERS TO OPEN QUESTIONS contains accurate single-sentence answers to the questions provided.

_____ _____

Signature Date

Appendix E

Human Grader Time Log

Instructions: Please enter the Student Name, Question Number and time spent grading each question in the spaces provided. After grading each test, provide a total of time spent (in hours, minutes and seconds). After grading all tests, provide a total of time spent (in hours, minutes and seconds).

Student Name	**Question Number**	**Time Spent Grading Question**
Total Time for Test		
Total Time for Test		
Total Time for Test		
Total Time for Test		
Total Time for Test		
Total Time		

Appendix F

Correct Answer and Student Responses in Canonical Form

Question	Key(shaded) & Student Responses	Canonical Form Generated
What is a cash-to-cash cycle?	A cash-to-cash cycle is the length of time it takes until a product has been manufactured from purchased materials.	time take product have be manufacture purchase material
Student A	A financial ratio showing for how long a company has to finance its own stock/inventory.	ratio show company have finance stock/inventory
Student B	A metric used to calculate how long cash is tied up in the main cash producing and cash consuming areas.	use calculate cash tie produce consume area
Student C	The cash-to-cash cycle calculates the time operating capital (cash) is out of reach for use by your business.	calculate time operate capital cash reach business
Student D	The number of days of working capital your organization has tied up in managing your supply chain.	number day work capital organization have tie manage chain
Student E	The length of time between the purchase of raw materials and the collection of accounts receivable generated in the sale of the final product.	time purchase material collection account generate sale product
What is encryption?	Encryption is the use of a mathematical formula that is applied to electronic data to render it illegible to anyone without a decoding key.	formula apply datum render key
Student A	Encryption is the activity of converting data or information into code.	activity convert datum information code
Student B	Encryption is the process of transforming information (referred to as plaintext) using an algorithm (called cipher) to make it unreadable to anyone except those possessing special knowledge, usually referred to as a key.	transform information referred plaintext use algorithm call cipher make possesses refer key
Student C	The process of converting ordinary language into code.	convert language code
Student D	To conceal information by means of a code or cipher.	conceal information mean code cipher

Question	Key(shaded) & Student Responses	Canonical Form Generated
Student E	A method of encoding data to prevent others from being able to interpret the information.	method encode datum prevent be interpret information
In supply chain management, what is a pull system?	A pull system is a supply chain in which the production of suppliers is determined by the needs of customers who request or order goods, necessitating production.	production supplier determine need customer good
Student A	The pull system enables the production of what is needed, based on a signal of what has just been sold.	enable production need base have be sell
Student B	A system where the production or movement of inventory items is initiated as required by the using department or location, or to replace items removed from an authorization queue.	production inventory item initiate require use department location replace remove authorization queue
Student C	The production of an item starts only when there is actual demand from a customer.	production item start demand customer
Student D	A supply system which requires that outlets request the amounts of commodities they need from higher-level storage facilities.	require outlet amount commodity nee storage facility
Student E	A system in which parts are only withdrawn after a request is made by the using operation for more parts.	part be withdraw make use operation
How can a log file be used for an eBusiness site?	A log file allows for its contents to be analyzed for information regarding the movements of users throughout a site.	allow content analyze information regard movement user
Student A	When a customer buys an item, his or her buying habits are logged into an extensive database and analyzed for potential future buying habits.	customer buy item buy habit be log database analyze
Student B	Log files will tell you which page your visitors were using when they decided to leave.	file tell page visitor be use decide leave
Student C	Log files count real people, when they visited, and whether they are newcomers or old-timers.	file count people visit be newcomer old-timer
Student D	Log files won't tell you how to fix a problem, but they'll let you know where the problem is.	file tell fix problem let know be

Question	Key(shaded) & Student Responses	Canonical Form Generated
Student E	Log files are records of all visitors to your site.	file be record visitor
Describe channel conflict.	Channel conflict is a situation in which the normal sales channels are bypassed.	be situation sale channel be bypass
Student A	Channel conflict occurs when manufacturers (brands) disintermediate their channel partners by selling their products direct to consumers through general marketing methods and/or over the internet through ecommerce.	occur manufacture brand disintermediate partner sell product consumer market method internet ecommerce
Student B	Discord in the channel.	Discord
Student C	Situation when a producer or supplier bypasses the normal channel of distribution and sells directly to the end user.	situation producer supplier bypass distribution sell user
Student D	Channel conflict is a situation in which channel partners have to compete against one another or the vendor's internal sales department.	be situation partner have compete vendor sale department
Student E	Channel conflict refers to a situation in which business partners clash in some of their operations in such a manner that it causes stress to the relationship.	refer situation business partner clash operation manner cause stress relationship