2010 10th IEEE International Conference on Advanced Learning Technologies

Automarking: Automatic Assessment of Open Questions

Laurie Cutrone and Maiga Chang
School of Computing and Information Systems, Athabasca University, Canada
lcutrone@rrc.mb.ca and maiga@ms2.hinet.net

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. In order to assess at higher levels of Bloom's (1956) taxonomy, 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. Automating the assessment process of open questions is an area of research that has been ongoing since the 1960s. Earlier work focused on statistical or probabilistic approaches based primarily on conceptual understanding. Recent gains in Natural Language Processing have resulted in a shift in the way in which free text can be evaluated. This has allowed for a more linguistic approach which focuses heavily on factual understanding. This study will leverage 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.**

Keywords-Natural Language Processing, Information Retrieval, WordNet, Part of Speech Tagging, Semantic Meaning, Open Question, Computerized Grading.

I. INTRODUCTION

Automating the assessment process of open questions is an area of research that has been ongoing since the 1960s [4][8][12]. However recent advances in Natural Language Processing, specifically in the areas of Information Extraction [2][21][24] and Information Retrieval [3][5][11][17] have allowed for alternative approaches to be explored. Natural Language Processing (NLP) uses computers to identify semantic relations among human words [9]. It involves various dimensions of human language including grammar, usage and semantics [16]. Earlier work in the area of Natural Language Processing, with respect to assessing responses to open questions, focused on a statistical or probabilistic approach [7][12][14][17]. 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 in general have resulted in a shift in the way in which free text can be evaluated. Specifically, work in the area of Information Extraction has allowed for the semantic meaning of natural language text to be captured [2][21][24]. This has allowed for a more linguistic approach which focuses heavily on factual understanding [1][23].

Much of the earlier work in the area of automatic essay grading was greatly influenced by an approach known as Latent Semantic Analysis (LSA) [7][22]. LSA uses a 'bag of words' approach in which similarity and co-location of words is evaluated [4]. 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 [7]. Two texts that use similar words would be considered semantically similar using LSA. 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 [4]. Therefore, the sentences: *The boy stepped on a spider*. And: *The spider stepped on a boy*. would be considered equivalent.

This study will leverage the recent work in Natural Language Processing to allow for a student response to be evaluated based on its *semantic* meaning. This is achieved through an extensive text pre-processing phase in which the semantic meaning of the response is captured. The architecture will incorporate the use of part of speech tagging as well as the WordNet database in an effort to alleviate the need for a large corpus in order to fairly assess the response. Additionally, 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.

This paper is organized as follows: Section II provides insight into the research areas that have contributed to this study. Section III provides an overview of the system architecture. Section IV describes the prototype system. Section V outlines the evaluation process for the system. Section VI provides areas in which this study could be enhanced in the future.

II. RESEARCH BACKGROUND

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

978-0-7695-4055-9/10 $26.00 © 2010 IEEE
DOI 10.1109/ICALT.2010.47

probabilities supported by corpus and the like to make a reasonable estimate at an appropriate grade.

Information Retrieval (IR) and Information Extraction (IE) are two sub-disciplines of Natural Language Processing. IR applies a model which specifies a process in which text may be compared with specific requirements to ultimately determine the relevance of the text [13]. IE involves the analysis of unrestricted text in an effort to extract relevant information. The relevant information extracted is based on some predefined guidelines [2]. Advances in both areas of Natural Language Processing will be 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 [17]. Chunking is the process of dividing sentences into noun phrases and verb groups [15]. 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 [17]. 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 [17].

IE has been used in an attempt to locate text that contains a predefined semantic meaning [21][24]. 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 [3][5]. 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 [6].

A significant offering to Natural Language Processing in recent years has been the development of WordNet[1] by George A. Miller of Princeton University. [18] 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 composed of synonymous words along with pointers to related synsets.

Part of Speech (POS) Tagging is a technique that has been widely used in Information Extraction Systems [6][10]. 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 [6][10].

III. System Architecture

This study will make use of the recent advances in Natural Language Processing to develop a system capable of automatically assessing open questions in a manner that assesses the student response based on its linguistic features. The system will reduce the student response as well as the supplied answer to their canonical form. All words in the canonical form will be tagged based on their part of speech. The student response and the supplied answer will then be compared. In this comparison, features encapsulated within WordNet will be utilized to ensure that exact word matches are not necessary in determining the level of equivalency between the student response and the supplied answer.

This system will utilize a component-based architecture. The components created in order to reduce the sentences to their canonical form will be used in the pre-processing of both the supplied correct answer as well as the student response. The basic architecture of the system is shown in Figure 1.

The system provides user interfaces for both the student and the assessor. In the Assessor User Interface, the assessor enters open questions in a free text editor using natural language. The assessor also provides natural language answers to the question. These answers represent the Correct Answer when responses are evaluated. The Student User Interface provides a student with a view of the open question(s) as well as a free text editor in which their response is collected. The student is to use natural language when providing a response.

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. These steps are applied to both the correct answer (CA) and the student response (SR). Figure 2 shows the effects of these steps on the sentence: *A cash-to-cash cycle is a measurement of the time it takes until resource inputs have been converted into cash flows*.

a. Text Tagging: This step involves providing each word and punctuation occurring in the sentence with a corresponding part of speech (POS) tag. This is accomplished using the SharpNLP[2] Part of Speech tagger. These tags are used extensively in various capacities throughout the system.

b. Removal of Punctuation: All punctuation is removed from the sentence. The reason for this is that punctuation is tagged differently than words using the SharpNLP Part of Speech tagger. Since the eventual canonical form of the sentence is no longer a correctly formed sentence, the absence of the punctuation does not impact the assessment process.

c. Removal of Question Words: Words found in the question should not be given credit when simply repeated in the student response. As a result all canonical words found in the question are removed from the correct answer as well as the student response.

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

[2] http://www.codeplex.com/sharpnlp

d. n-Gram Detection/Transformation: It is important to recognize word groupings that connote a single meaning. These include compound words or proper nouns [20]. For example, the word grouping "telephone directory" should not be split even though it is comprised of individual nouns that, in and of themselves, connote meaning. This pre-processing step will re-tag any identified n-Grams.

e. Reverse Context: Natural language text can have a variety of morpho-syntactic variations which are equivalent semantically [19]. In some cases, a sentence can be stated in a reverse form which is equivalent to a more direct approach. 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 tags identified in the Text Tagging step will be utilized in this step.

f. Stop Word Processing: In this step, the tagged text is examined to locate and remove stop words. This will cause most sentences to be grammatically incorrect. However, the semantic meaning of the sentence remains.

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 [10]. 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 will simplify the process of locating synonyms which takes place following the pre-processing phase.

Following text pre-processing, the evaluation of the student responses based on the correct answer takes place. The Correct Answer in Canonical Form (CFCA), as well as the Student Response in Canonical Form (CFSR) are initially compared to look for exact word matches. Any unmatched words are then applied to WordNet. In this step, all synonyms for all unmatched words in the CFCA and CFSR are determined. For matched words as well as matched synonyms, the part of speech tag as well as the relative position of the words are evaluated to determine whether the match is accurate. The final step in the matching algorithm is to apply a value between 0 and 1 to each remaining word in the supplied answer. An exact match with appropriate relative positions to the other words as well as matching part of speech tags would be assigned a value of 1. Synonymous matches with an appropriate relative position and matching part of speech tags would be given a value based on its level of equivalence within the WordNet web.

The final grade of the open question is calculated by applying a formula that considers the value assigned to the remaining words in the supplied answer as well as the weight of the question. At this point each remaining word in the supplied answer is given an equal share of the overall weight of the question.

IV. Prototype System

A prototype system has been developed using the architecture described above and shown in Figure 1. The prototype system contains the user interfaces described in the System Architecture section above. Both the correct answer and the student response are processed using the text pre-processing steps in order to reduce these sentences to their canonical forms. The canonical forms are then compared to one another to determine their level of equivalency. This comparison incorporates all synonyms of the canonical words in order to allow for more flexibility in terms of the choice of words in the correct answer as well as the student response. The degree to which similarity of words is determined is based on the distance between words within the WordNet hierarchy. Essentially, the greater the distance, the less related two words are.

Early testing of the prototype system has produced some encouraging results. For example, the sentence: *The boy stepped on a spider*, and: *A spider was stepped on by the boy*, are considered equivalent. While a third sentence: *A spider stepped on the boy* is not considered equivalent to the other two. This comparison is successful as a result of the reverse context component in the text pre-processing phase which looks for 'was/by' phrasing, and simplifies the sentence by reversing the text and removing the 'was/by' phrasing. Additionally, the third sentence is considered incorrect based on the placement of the nouns (spider and boy) with respect to the verb (stepped) even though it contains all of the words found in the first sentence.

Another successful test has occurred as a result of the question: *What is a cash-to-cash cycle?* The supplied correct answer is: *A cash-to-cash cycle is a measurement of the time it takes until resource inputs have been converted into cash flows*. The preprocessing of this sentence is shown in Figure 2, which results in the eventual canonical form: *measure time convert resource input cash flow*.

V. Evaluation Plan

The system will be evaluated using questions within the E-Commerce domain. Questions within this domain will be entered into a search engine to produce multiple differing responses of varying levels of accuracy and correctness. These varied answers will be submitted to the system to simulate varied student responses. The system will then process the responses and assign a grade to each response. The same responses will be graded by two independent human graders in order to determine the level of agreement between the human graders and the Automarking system. This agreement level will be also contrasted with the level of agreement between the two human graders.

The above method of evaluation does not consider the breadth or depth of knowledge expected by the simulated students. For example, the supplied correct answer may be non-technical in nature and a simulated student response may be correct, but a much more technical response. As a result, the student response may be mis-graded. With this in mind, a follow-up evaluation is recommended in which actual students who have had the same instruction and who are working towards the same learning outcomes would provide responses to the questions.

Early testing of the individual components of the system has been promising. Sentences with similar meaning have been given appropriate grade values. The English language is a complex language. For every language rule, there are countless exceptions to that rule. As a result it is necessary that further, extensive testing take place in order to provide a complete evaluation of the system.

VI. Future Work

This study was developed under very strict constraints. The system in its current format is capable of processing answers containing a single sentence that is free of grammar and spelling mistakes. Future work is encouraged which would allow for multiple sentences to be graded based on their collective meaning. Additionally, future work could incorporate a spell checker and grammar checker. Future work should also allow for a more flexible marking algorithm in which the canonical words could be given varying weight values in the grading scheme depending on their level of importance in the student response.

Acknowledgment

The authors wish to thank the support of Athabasca University and the Mission Critical Research funding. Additional acknowledgements go out to the Natural Language Processing community. This growing community provides a great deal of support to developers working on leading edge applications of Natural Language Processing. This support takes on many forms such as active forums, robust and well-supported freeware, along with the growing number of Natural Language Processing experts.

References

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

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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).

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

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

[18] 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.

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

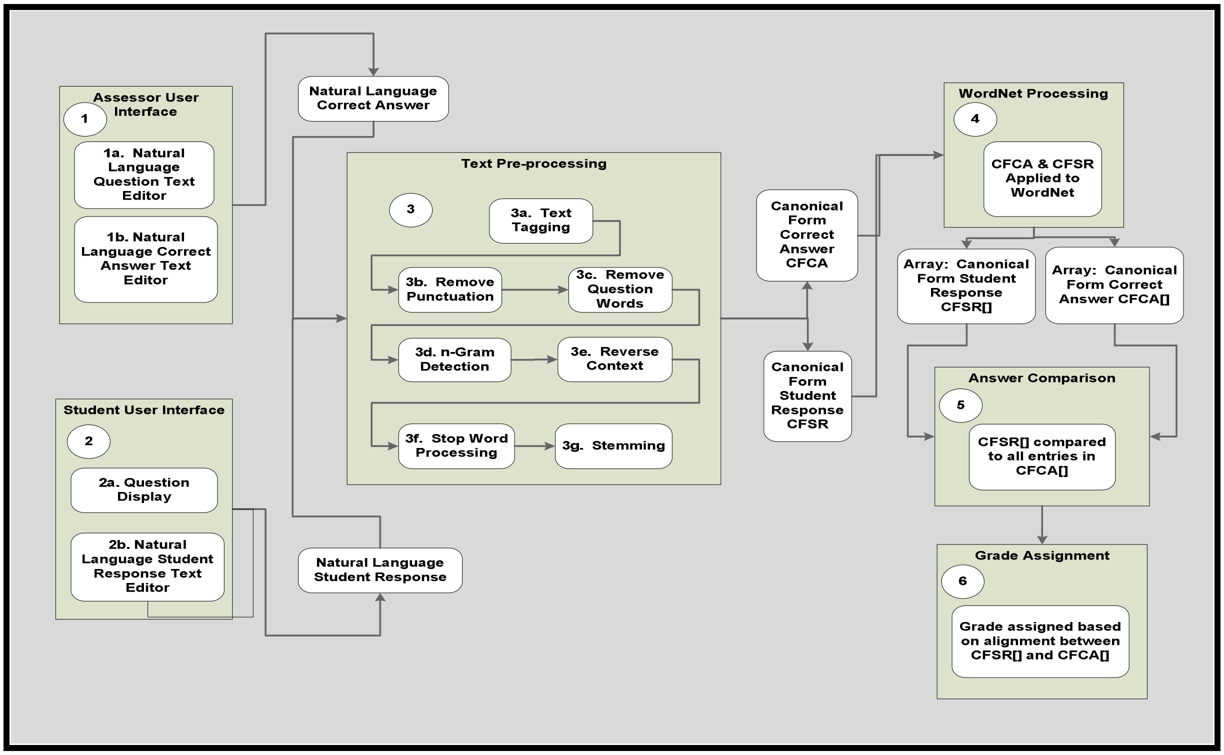
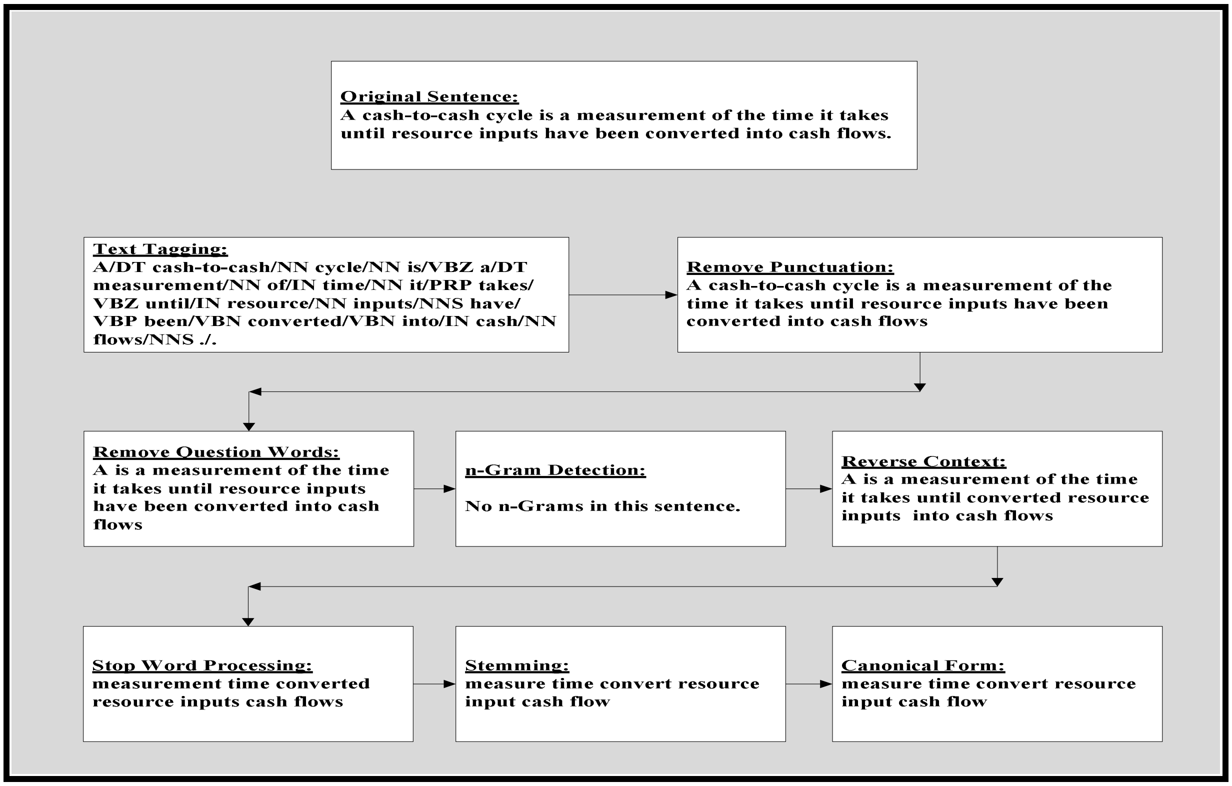

Assessor User Interface
1
1a. Natural Language Question Text Editor
1b. Natural Language Correct Answer Text Editor
Natural Language Correct Answer
Text Pre-processing
3
3a. Text Tagging
3b. Remove Punctuation
3c. Remove Question Words
3d. n-Gram Detection
3e. Reverse Context
3f. Stop Word Processing
3g. Stemming
Canonical Form Correct Answer CFCA
Canonical Form Student Response CFSR
WordNet Processing
4
CFCA & CFSR Applied to WordNet
Array: Canonical Form Student Response CFSR[]
Array: Canonical Form Correct Answer CFCA[]
Answer Comparison
5
CFSR[] compared to all entries in CFCA[]
Grade Assignment
6
Grade assigned based on alignment between CFSR[] and CFCA[]
Student User Interface
2
2a. Question Display
2b. Natural Language Student Response Text Editor
Natural Language Student Response

Figure 1 - System Architecture



Original Sentence:
A cash-to-cash cycle is a measurement of the time it takes until resource inputs have been converted into cash flows.

Text Tagging:
A/DT cash-to-cash/NN cycle/NN is/VBZ a/DT measurement/NN of/IN time/NN it/PRP takes/VBZ until/IN resource/NN inputs/NNS have/VBP been/VBN converted/VBN into/IN cash/NN flows/NNS ./.

Remove Punctuation:
A cash-to-cash cycle is a measurement of the time it takes until resource inputs have been converted into cash flows

Remove Question Words:
A is a measurement of the time it takes until resource inputs have been converted into cash flows

n-Gram Detection:
No n-Grams in this sentence.

Reverse Context:
A is a measurement of the time it takes until converted resource inputs into cash flows

Stop Word Processing:
measurement time converted resource inputs cash flows

Stemming:
measure time convert resource input cash flow

Canonical Form:
measure time convert resource input cash flow

Figure 2 - Sentence Evolution