Hybrid Ontology-based Information Extraction for Automated Text Grading

Fernando Gutierrez*, Dejing Dou*, Adam Martini*, Stephen Fickas* and Hui Zong†
* Department of Computer and Information Science
University of Oregon, Eugene, OR 97403
Email: {fernando, dou, martini, fickas}@cs.uoregon.edu
† Department of Microbiology,
University of Virginia, Charlottesville, VA 22093
Email: hz9s@virginia.edu

***Abstract*—Although automatic text grading systems have reached an accuracy level comparable to human grading, with successful commercial and research implementations (e.g., Latent Semantic Analysis), these systems can provide limited feedback about which statements of the text are incorrect and why they are incorrect. In the present work, we propose the use of a hybrid Ontology-based Information Extraction (OBIE) system to identify both correct and incorrect statements by combining extraction rules and machine learning based information extractors. Experiments show that given 77 student answers to a Cell Biology final exam question, our hybrid system can identify both correct and incorrect statements with high precision and recall measures.**

I. Introduction

With the help of advances in Natural Language Processing, automatic text grading has shifted from simple sequence comparison into systems that capture the underlying semantics of the text, with accuracy comparable to a human grader. However, since most methods verify correctness by measuring similarity between student writings and a standard text, limited feedback can be given regarding errors. Following a different approach, Brent et al. [1] propose the use of domain knowledge in the process of text grading to help determine the correctness of the text. Similarly, in our previous work [2] we incorporate formal knowledge representation, through ontology, into the process of evaluating the correctness of text.

In this paper, we extend our previous work by presenting a hybrid Ontology-based Information Extraction (OBIE) system for extracting both correct and incorrect statements in automatic text grading. We have identified that the information extractors, which are the OBIE components that do the extraction process, can have multiple dimensions. These dimensions allow the information extractors to represent different ontological concepts, to have different implementations (e.g., machine learning and extraction rules), and to perform different functions (i.e., extracting correct or incorrect statements [2]). These dimensions allow us to combine different information extractors in one hybrid OBIE system, letting the system have multiple configurations. We argue that the combination of information extractors that perform different *functions* can provide a better understanding of a graded text, and the combination of information extractors that have different *implementations* can improve the performance of the extraction process.

We have applied our hybrid OBIE system to identify the correct and incorrect statements of 77 student answers from an undergraduate biology class. We have also applied our system to a set of synthetic datasets to determine the scalability of our system given a large dataset. We have found that the combination of information extractors that have different implementations can obtain a higher precision and recall than using only one type of implementation. We have also found that the extraction of incorrect statements is more complex than the extraction of correct statements, leading to a higher variability in the performance of information extractors. The experiment results show that performance variability can be reduced through the use of a hybrid configuration.

II. Related Work

Automatic text grading systems have been studied for many years [3]. Mitchell et al. [4] have identified three main approaches to solve this problem. The first approach is *keyword analysis*. In this approach, grading is based on the identification of coincident words and n-grams, and it includes machine translation techniques [5] and n-gram co-occurrence methods [6]. The second approach uses Natural Language Processing (NLP), which tries to capture meaning through semantic analysis or deep parsing. The most popular NLP approach is Latent Semantic Analysis (LSA) [7], [8], which attempts to identify the underlying semantics of the text through Singular Value Decomposition (SVD). Finally, the third approach is based on Information Extraction (IE) [1], [4]. Without doing deep analysis, IE uses NLP tools to identify specific information from plain text.

Most of the previously mentioned methods are highly accurate when compared to human grader (e.g., LSA). Yet, because these systems' evaluation is only based on the correct elements of the text or on its similarity to a *standard* essay/summary, the feedback they can provide does not consider indications of what is missing or what is incorrect in the students' text. In order to overcome this limitation, SAGrader [1] incorporates domain knowledge through the use of *semantic network*. This representation allows SAGrader to evaluate student's writings based on the correctness and completeness (missing concepts and relations) of the text.

Similarly to SAGrader, in our previous work [2], we propose the use of Ontology-based Information Extraction

978-0-7695-5144-9/13 $31.00 © 2013 IEEE
DOI 10.1109/ICMLA.2013.73

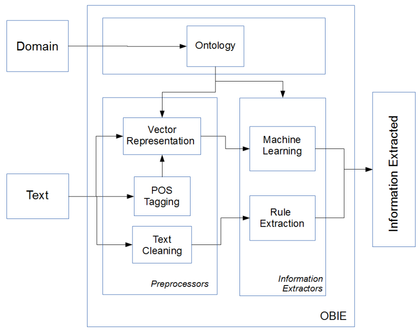

Fig. 1. A Hybrid Ontology Based Information Extraction System

(OBIE) for automatic summary grading. In OBIE, an ontology provides a formal representation of the domain, which guides the information extraction process [9]. Ontologies can provide richer representations than semantic networks by offering consistency (e.g., semantic networks have ambiguous interpretations of ISA relationships) [10], and can represent disjointness and negations [11]. By incorporating a heuristic ontology debugging technique [12] into our OBIE approach, we can determine axioms that can create logical contradictions in the domain. These axioms are translated into rule-based information extractors that identify incorrect statements. By knowing which statements are inconsistent with respect to the ontology and why (through the inconsistent axioms), it is possible to produce more detailed and accurate feedback.

III. A Hybrid Ontology-based Information Extraction System

In the present work, we propose a hybrid Ontology-based Information Extraction (OBIE) system for extracting correct and incorrect statements from students writings. The system integrates information extractors of different ontological concepts and relationships, different implementations (i.e., manually written extraction rules and machine learning generated information extractors), and different functions (i.e., extracting correct statements and extracting incorrect statements). This is an important extension to Ontology-based Components for Information Extraction (OBCIE) designed by Wimalasuriya and Dou in 2010 [13]. The main goal behind OBCIE is to promote re-usability, so the components are defined as modular as possible. This modularity becomes more obvious when specifying information extractors according to different ontological concepts and relationships, different implementations, and different functions. This approach allows the possibility of combining different sets of extractors, and integrating into the extraction system only those that are required or relevant to the task.

For the present work, the following major *OBCIE components* have been selected to construct a hybrid OBIE system (Figure 1):

1) Ontology: Provides formal representation of the concepts and relationships of a domain.
2) Preprocessors: These convert the text into a format that can be processed by components of the next phase, information extractors.
3) Information Extractors: These make extractions for both correct and incorrect information with respect to a specific class or a property of an ontology.

In the following sections we give details on how each of the mentioned modules is implemented.

A. Ontology

Given a domain, there may be a well-developed ontology (e.g., Gene Ontology [14]). We can directly adopt the partial or the whole ontology depending on how much domain knowledge we need for the information extraction. If there is no well-developed ontology, we can design a new application driven ontology which only covers the knowledge needed for the information extraction.

B. Preprocessing

The type of preprocessing used in an OBIE system depend on the type of information extractors being used by the system. In the present work, we define two sets of preprocessing processes:

- **Preprocessing for Machine Learning:** In most cases, machine learning techniques require the text to be transformed, usually into some numerical representation. For the classifier, the text is represented as a binary vector. The vector also contains *part-of-speech* information (e.g., number of verbs and nouns) as features. For the probabilistic model, the text is enhanced with *part-of-speech* tags, output of the first stage, and some extra features proposed by Wu and Weld [15] for the *Kylin* system.
- **Preprocessing for Extraction Rules:** For the extraction rules in our system, the preprocessing consists of cleaning the non-letter characters from text (e.g., removing special end-of-line characters) and a limited amount of *text merging* (e.g., "ETCs" and "electron transport chain" are replaced by "ETC").

C. Information Extractors

In general, the process of extracting information from plain text can be done by applying *extraction rules* or by using *machine learning* generated information extractors [9]. Based on regular expression, extraction rules capture information by identify specific elements in text. In most cases, extraction rules are simple to design, and they have relatively good performance. However, because they are based on specific cues crafted manually, extraction rules are difficult to generalize and do not scale well.

On the other hand, with machine learning methods such as Support Vector Machine, Naive Bayes, or Conditional Random Fields, the information extraction task is transformed into a labeling and supervised learning task, where classification methods and probabilistic models try to identify which elements from a sentence are part of the sought information. These techniques obtain good accuracy, and they scale well. However, machine learning techniques are data-driven, so the performance of these methods depend on the quality and quantity of the data used for the training.

As consequence of these strengths and weaknesses, some ontological concepts are more difficult to extract than others for any given approach. In order to maximize the extraction capabilities of our system, we have incorporated both extraction rules and machine learning based information extractors into our hybrid OBIE system.

1) Extraction Rules as Information Extractors: As previously mentioned, extraction rules are based on regular expression, and they can capture different types of elements from text. Although it is possible to combine multiple extraction rules with different level of abstraction (i.e., hierarchy of extraction rules) for extracting one concept, in the present work we define single, one level extraction rules. Each extraction rule represents one axiom of the ontology. Because we have an additional goal to identify statements that are inconsistent with the ontology, we have defined extraction rules for the consistent and inconsistent axioms of each concept.

2) Machine Learning Based Information Extractors: The machine learning based information extractors are implemented following a two-phase classification scheme. In the first phase, the method identifies which sentences from the document contain the information the extractor seeks. The process is defined as a binary classification task (Naive Bayes), where one class corresponds to sentences that carry the information and the other class corresponds to sentences that do not have the information. The text is transformed into a binary vector, which also contains some metadata similar to the Kylin system [15].

The second phase of the platform identifies the elements of the sentence (words) that contain the information. This is done by a probabilistic model (Conditional Random Fields). For this phase the text is enhanced with part-of-speech labels, metadata information used in the first phase, the output of the previous phase, and a group of extra features that are proposed and used by the Kylin system [15].

D. Functions of the Information Extractors

We have designed two functions for both extraction rules and machine learning based information extractors:

1) Extracting Correct Statements: Under the intuition that an answer to a domain specific question must be entailed from the domain, we consider a statement to be correct if it is consistent with the domain's knowledge. In other words, in the case of extracting correct statements, we use axioms from the ontology to design the information extractors.

2) Extracting Incorrect Statements: Because the answer is a logical consequence (entailment) of the concepts and relationships of the domain ontology, it is natural to consider that an incorrect answer will be inconsistent with the domain ontology. If we can identify the axioms that are inconsistent to each other, then we can detect incorrect statements given a consistent ontology. Therefore, research in ontology inconsistency provides some insight in how to detect incorrect statements.

Based on their observations from workshops and tutorials related to ontology construction, Wang et al. [12] have proposed a heuristic approach to identify the cause of inconsistency in an ontology. Following Wang et al.'s approach, it is possible to determine a set of *incorrect axioms* that would

$$\frac{\begin{array}{c}\forall x \forall y \Big(\big(Professor(x) \wedge Teaches(x,y) \big) \rightarrow Student(y) \Big) \\ \forall x \Big(Professor(x) \leftrightarrow \neg Student(x) \Big)\end{array}}{\forall x \forall y \Big(\big(Professor(x) \wedge Teaches(x,y) \wedge Professor(y) \big) \rightarrow \perp \Big)}$$

TABLE I. EXAMPLE OF AXIOM FOR EXTRACTING INCORRECT STATEMENTS.

make the ontology inconsistent. These *incorrect axioms* define the information extractors for detecting incorrect statements. For example, axiom in Table I means if an ontology defines 1) a person who is taught by a professor must be a student, 2) a professor cannot be a student, then any statement about a professor teach another professor is incorrect.

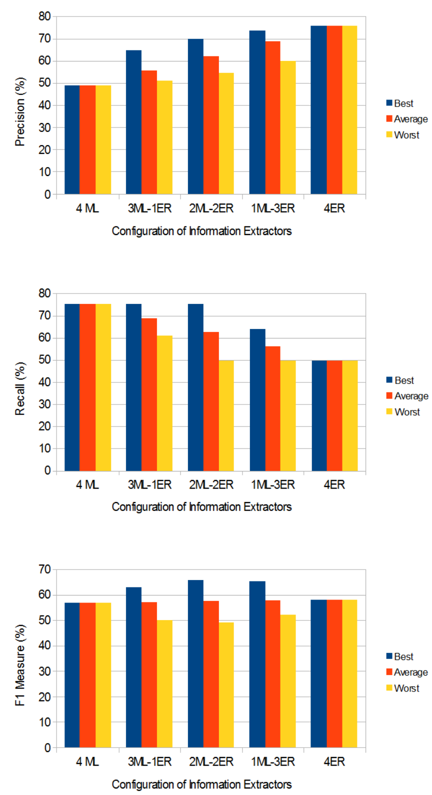

Fig. 2. Precision, recall and F1 measure for different configurations of machine learning (ML) extractors and extraction rules (ER) when applied to the *real dataset*

IV. EXPERIMENTS

As previously mentioned, our proposed hybrid OBIE system allows the combination of information extractors. These extractors are from different ontological concepts and relationships, might have different implementations, and that perform different functions. To evaluate our proposal we have considered two sets of experiments: the first set of experiments use data collected from an undergraduate biology class (*real dataset*), while the second set of experiments use a *synthetic data* set generated from correct and incorrect statements to test the scalability of our system.

Element type	Number of element
Concepts	17
Relationships	10
Subclass relationships	3

TABLE II. STATISTICAL INFORMATION ABOUT THE ONTOLOGY.

A. Real Dataset

The *real dataset* correspond to student answers in the final exam of an undergraduate biology class. The corpus consists of 77 student answers. From the exam, we selected one question which requires the students to present a justified answer: *"If you generate a mutation that breaks down the electron transfer chain in mitochondria, will myosin proteins fall off microfilaments or get stuck to it? Why?"*

Each answer is a short paragraph that may contain at most four sentences. The answers have been labeled by domain experts (the instructor of the class and his teaching assistants) indicating if they are correct or incorrect, and if the answers provide enough justification. An example of a correct answer is:*"They will tend to get stuck, because the exchange of ATP for ADP causes the myosin head to release the microfilament. If the ETC is halted, ATP will no longer be produced."* An example of an incorrect answer is: *"They will fall off. This is because a mutation in the ETC will cause an absence of ATP."*

1) Ontology: Although there are many biology-related ontologies available (the National Center of Biomedical Ontology's BioPortal[1] website offers access to more than 300 biomedical ontologies), they do not offer the necessary relationships that are required to analyze the students' answers. To overcome this limitation we have developed our own ontology.

To construct the ontology we have followed two main guidelines: it must contain all concepts and relationships that will allow answering the exam's question, and it must not include any other concepts that are not required to answer the question. The first requirement intends to provide the sufficient domain knowledge to analyze the arguments of the answer, i.e., why the *myosin* is affected by *mitochondrial* defect. The second requirement tries to reduce the complexity of the ontology by keeping it focus on the part of the domain that is relevant for the task. This criteria leads to an ontology that is highly connected, but has a small number of hierarchical relationships between concepts (see Table II).

2) Information Extractor Creation and Evaluation: We have created information extractors considering the three dimensions previously mentioned. From the ontological perspective, we have created extractors for three concepts *Myosin*, *ETC*, and *ATP*. However, given the labels and distribution of statements, only the concept *Myosin* has extractors for both correct and incorrect statements. This leads to four types of extractors: *correct Myosin*, *incorrect Myosin*, *correct ETC*, and *correct ATP*.

We have implemented each one of the four types of extractors as machine learning generated extractors and manually generated extraction rules. In the case of machine learning, 50% of the corpus was used for training each extractor and the rest was used for testing. Similar way for the extraction rules, 50% of the corpus was used to create the extraction rule of each extractor, and the rest was used for testing it.

3) Results: Considering that there are four types of concepts and two types of implementations, we have identified five possible configurations of information extractors that our hybrid OBIE system can use. There are two *pure* configurations: using all four machine learning extractors (4ML), or using all four extraction rules (4ER). There also are three *hybrid* configurations: using three machine learning extractor and one extraction rule (3ML-1ER), using two machine learning extractors and two extraction rules (2ML-2ER), and using one machine learning extractor with three extraction rule extractors (1ML-3ER).

The *pure* configurations have a unique setting, in which all the information extractors use the same type of implementation. However, in the case of the *hybrid* configurations, each configuration allows multiple types of settings. For example, in the case of using three machine learning extractors and one extraction rule (3ML-1ER), we can choose an extraction rule implementation for any one of the four concepts and use machine learning extractors for the rest. There are four or six possible settings per *hybrid* configuration.

In order to provide clarity, we summarize our results by presenting the average *precision* (correctness of the extraction), average *recall* (completeness of the extraction), and average *F1* measure (overall performance of the extraction) for each configuration. Each one of these measurements represents the average performance of all possible settings of a given configuration. In turn, the performance of a setting is the average performance of all information extractors of that setting. We also report the best and worst performances of each configuration. In the case of the pure configurations, the average, best, and worst performances are the same because they only have one setting. The precision, recall and F1 measure for the real dataset are presented in Figure 2.

The results show that extraction rules have a higher precision than machine learning based extractors, and when a *hybrid* configuration has more extraction rule extractors, it has a higher performance in precision. On the other hand, machine learning extractors have higher recall than extraction rules. This effect can also be seen in the *hybrid* configurations with more machine learning extractors. For precision and recall, the *hybrid* configurations perform within the range defined by the *pure* configurations. This leads to *hybrid* configurations having a higher F1 measure than the *pure* configurations.

B. Synthetic Dataset

Because of the limitations presented by the *real dataset*, we have created a *synthetic dataset* to explore scalability issues.

1) Data Generation: As previously mentioned, the correct answer for the exam's question can be constructed by combing statements from four concepts (*Myosin*, *ETC*, *ATP*, *ADP*). Therefore, to generate an answer, we need to create a paragraph that contains a statement from each concept.

Based on the ontology and some statements from the students answers, we have constructed sets of correct and incorrect statements for each of the four concepts. In general, the sets of incorrect statements are much larger than the sets of

[1] http://bioportal.bioontology.org/

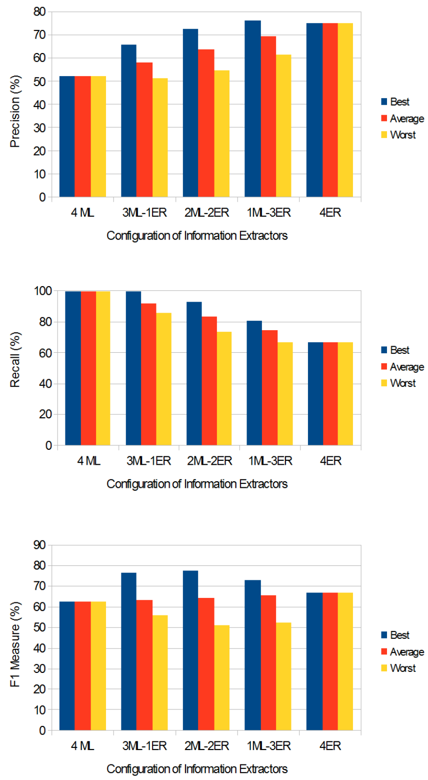

Fig. 3. Precision, recall and F1 measure for different configuration of machine learning (ML) extractors and extraction rules (ER) when applied to the synthetic datasets (correct statement extractors).

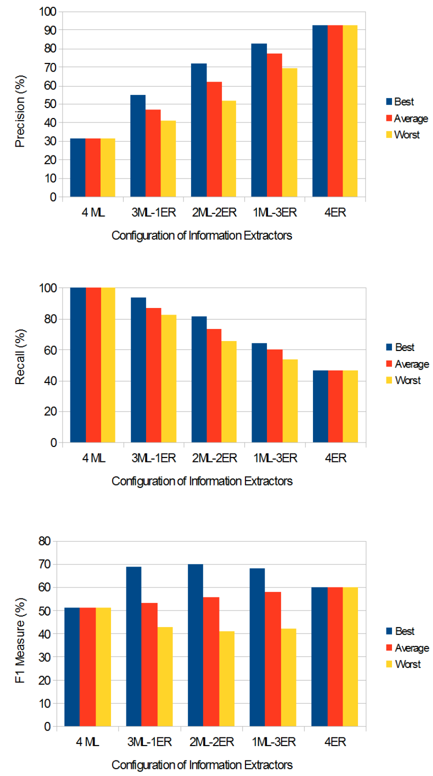

Fig. 4. Precision, recall and F1 measure for different configuration of machine learning (ML) extractors and extraction rules (ER) when applied to the *synthetic datasets* (incorrect statement extractors).

correct statements because the incorrectness of a statement can be caused by more than one reason. To generate an instance (an answer), we randomly select one statement from each concept. And given the level of errors of the synthetic dataset, an incorrect statement might be selected. We created datasets containing 1000 instances with error levels of 20%, 25%, 30%, 40%, 50%, 60%, 70%, 75%, 80%, and 90%.

2) Information Extractor Creation and Evaluation: We have created 16 information extractors by combining all possible four concepts (*Myosin*, *ETC*, *ATP*, *ADP*), two implementations (i.e., machine learning and extraction rules), and two functions (i.e., extracting correct and incorrect statements).

The machine learning extractors are constructed and evaluated by 10-fold cross validation. In other words, the dataset is randomly sorted and divided into 10 folds, nine folds are used in training and the non-selected fold is used for testing. For the extraction rules, we have selected 20% of the instances of a dataset for development of extraction rules, while the rest 80% is used for evaluation.

3) Results: In order to provide better understanding of the effect of the information extractor's implementations over its functions (i.e., extracting correct and incorrect statements), we have tested our system in a similar way as with the *real dataset*. We have identified five configurations of information extractors for each function: using four machine learning extractors (4ML), using four extraction rules (4ER), using three machine learning extractor and one extraction rule (3ML-1ER), and using two machine learning extractors and two extraction rules (2ML-2ER), and using one machine learning extractor with three extraction rule extractors (1ML-3ER).

In the case of correct information extractions, the precision of the pure configurations (4ML and 4ER) seem to have a similar behavior as the extractors of the *real dataset*, with extraction rules obtaining higher precision (Figure 3). On the other hand, although the mix configurations follow the trend of higher precision when more extraction rules are used, we can observe configurations that obtain lower precision (3ML-1ER) and higher precision (1ML-3ER) than the pure configurations. For the recall of the correct information extractors, we can observe that machine learning generated information extractors outperform extraction rules by a larger margin, and that effect can also be observed in the performance of the mix configurations. The F1 measure indicates that the overall average performance of all configurations are similar, with a slight trend of higher F1 for extraction rule implementations. The best and worst F1 for the mix configurations do show a difference in performance.

In general, incorrect information extractors perform in similarly to correct information extractors, with extraction rule implementations obtaining higher precision and machine learning implementation showing better recall. Looking more closely at the results, we see that the difference between best

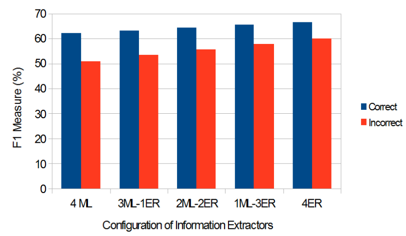

Fig. 5. F1 measure of different configuration of machine learning (ML) extractors and extraction rules (ER) when applied to the synthetic datasets. Comparison between information extractors for correct and incorrect statements.

and worst performances in both precision and recall increased drastically. This difference also appears in F1 measure (Figure 4).

Finally, Figure 5 compares the average performance of each configuration given their functionality. In general, all configurations for correct information extractors have a higher F1 measure than their incorrect counterpart. Figure 5 also shows that extraction rules have a slightly better performance than machine learning, with this tendency being more clear for the incorrect information extractors.

V. Conclusions

We have presented a hybrid Ontology-based Information Extraction system in the application of automatic text grading. We demonstrate how these information extractors can perform different functions and have different implementations. By changing the settings of information extractors, the proposed OBIE system can identify both correct and incorrect statements to provide a better understanding of the graded text. We also show how different implementations can produce optimized performance.

We have applied our hybrid OBIE system to identify the correct and incorrect statements of 77 student answers from an undergraduate biology class, evaluating multiple configurations of information extractors. We have also applied our system to a set of synthetic datasets with different levels of errors to test how our system scales to larger datasets. Our hybrid system can identify both correct and incorrect statements with high and balanced precision and recall measures. In more detail, we have found that the combination of information extractors that have different implementations can obtain a higher precision and recall than using only one type of implementation. We also found that the extraction of incorrect statements is more complex than the extraction of correct statements, which leads to high variability in the performance of information extractors. The experiment results show that this variability can be reduced through the use of a hybrid configuration.

In terms of future work, we intend to incorporate automatic configuration selection of information extractors, based on the results obtained in this work. And based on results obtained by Wimalasuriya and Dou [16], it seems possible to aggregate the output of information extractors from multiple and heterogeneous ontologies.

VI. Acknowledgments

This research is partially supported by the National Science Foundation grant IIS-1118050 and grant IIS-1013054. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the NSF.

References

[1] E. Brent, C. Atkisson, and N. Green, *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-Learning Support*. IGI Global, 2010, ch. Time-Shifted Online Collaboration: Creating Teachable Moments Through Automated Grading, pp. 55–73.

[2] F. Gutierrez, D. Dou, S. Fickas, and G. Griffiths, "Providing Grades and Feedback for Student Summaries by Ontology-based Information Extraction," in *Proceedings of the 21st ACM Conference on Information and Knowledge Management (CIKM)*, 2012, pp. 1722–1726.

[3] E. B. Page, "The imminence of... grading essays by computer," *The Phi Delta Kappan*, vol. 47, no. 5, pp. pp. 238–243, 1966. [Online]. Available: http://www.jstor.org/stable/20371545

[4] T. Mitchell, T. Russell, P. Broomhead, and N. Aldridge, "Towards robust computerised marking of free-text responses," in *Proceedings of the 6th CAA Conference*. Loughborough University, 2002. [Online]. Available: http://hdl.handle.net/2134/1884

[5] D. Perez, E. Alfonseca, and P. Rodriguez, "Upper bounds of the bleu algorithm applied to assessing student essays," in *Proceedings of the 30th International Association for Educational Assessment (IAEA) Conference*, 2004. [Online]. Available: http://orestes.ii.uam.es:8080/dperez/articulos/EnrDiaPilIAEA04.pdf

[6] C.-Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Workshop On Text Summarization Branches Out*, 2004, pp. 25–26.

[7] M. Franzke and L. Streeter, "Building student summarization, writing and reading comprehension skills with guided practice and automated feedback," White paper from Pearson Knowledge Technologies, 2006.

[8] Y. He, S. C. Hui, and T. T. Quan, "Automatic summary assessment for intelligent tutoring systems," *Computers and Education*, vol. 53, no. 3, pp. 890–899, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.compedu.2009.05.008

[9] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *J. Inf. Sci.*, vol. 36, pp. 306–323, June 2010. [Online]. Available: http://dx.doi.org/10.1177/0165551509360123

[10] D. McDermott, "Artificial intelligence meets natural stupidity," *SIGART Newsletter*, no. 57, pp. 4–9, April 1976. [Online]. Available: http://doi.acm.org/10.1145/1045339.1045340

[11] S. J. Russell and P. Norvig, *Artificial intelligence : a modern approach.*, 3rd ed. Prentice Hall, 2010, ch. Knowledge representation, p. 454.

[12] H. Wang, M. Horridge, A. Rector, N. Drummond, and J. Seidenberg, "Debugging owl-dl ontologies: A heuristic approach," in *Proceedings of the 4th international conference on The Semantic Web*. Springer, 2005, pp. 745–757.

[13] D. C. Wimalasuriya and D. Dou, "Components for Information Extraction: Ontology-Based Information Extractors and Generic Platforms," in *Proceedings of the 19th ACM Conference on Information and Knowledge Management (CIKM)*, 2010, pp. 9–18.

[14] T. Gene Ontology Consortium, "Creating the Gene Ontology Resource: Design and Implementation," *Genome Research*, vol. 11, no. 8, pp. 1425–1433, 2001.

[15] F. Wu and D. S. Weld, "Autonomously semantifying wikipedia," in *CIKM*. NY, USA: ACM, 2007, pp. 41–50.

[16] D. C. Wimalasuriya and D. Dou, "Using Multiple Ontologies in Information Extraction," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, 2009, pp. 235–244.