Providing Grades and Feedback for Student Summaries by Ontology-based Information Extraction

Fernando Gutierrez
Computer and Information Science Department
University of Oregon, USA
fernando@cs.uoregon.edu

Dejing Dou
Computer and Information Science Department
University of Oregon, USA
dou@cs.uoregon.edu

Stephen Fickas
Computer and Information Science Department
University of Oregon, USA
fickas@cs.uoregon.edu

Gina Griffiths
Special Education and Clinical Sciences Department
University of Oregon, USA
ginag@uoregon.edu

ABSTRACT

Automatic grading systems for summaries and essays have been studied for years. Most commercial and research implementations are based in statistical methods, such as Latent Semantic Analysis (LSA), which can provide high accuracy on similarity between the essay and the graded or standard essays, but they can offer very limited feedback. In the present work, we propose a novel method to provide both grades and meaningful feedback for student summaries by Ontology-based Information Extraction (OBIE). We use ontological concepts and relationships to create extraction rules to identify correct statements. Based on ontology constraints (e.g., disjointness between concepts), we define patterns that are logically inconsistent with the ontology to create rules to extract incorrect statements. Experiments show that the grades given to 18 student summaries on Ecosystems by OBIE are correlated to human gradings. OBIE also provide meaningful feedback on the errors those students made in their summaries.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous—*Experimentation*

Keywords

Automatic Grading, Information Extraction, Ontology

1. INTRODUCTION

A reading-comprehension strategy that forces a student to retrieve or recall information from memory has a potent effect on learning, enhancing long-term retention of the tested information [5]. Researchers have found this not only to be true when external tests are administered (e.g., a midterm), but also when students do self-testing, e.g., attempt to write summaries on a section of reading [7]. However, it is important that the self-test is analyzed, errors are found, and feedback is delivered to the student. Using this feedback, the student is expected to review material with a goal of correcting errors in comprehension.

Because of the advances in Natural Language Processing (NLP), automatic grading of summaries and essays has become possible, with several commercial applications [1, 3]. Most existing automated grading systems for student summaries are based on statistical models, such as Latent Semantic Analysis (LSA) [6]. Although the statistical NLP based systems produce quite accurate grades, they cannot provide feedback about the completeness or correctness of the summaries, especially what errors the students have made in their summaries [1].

In this paper, we present a new approach for automatic summary grading and error detection, which uses Ontology-based Information Extraction (OBIE). The objective of OBIE is to use a domain ontology, which specifies the concepts and relationships for a particular domain, to guide the process of Information Extraction [13]. We have applied OBIE to automatic grading and error detection for 18 student summaries in the domain of Ecosystems, extracting both logically correct and incorrect sentences from the summaries. We evaluated the performance of the OBIE-based grading system according to three measures, and compared it against the grades provided by Latent Semantic Analysis (LSA) and an expert human grader on the same summaries.

The rest of this paper is organized as follows. We introduce some related work in Section 2 while presenting our OBIE method and implementation in Section 3. We report our experimental results in Section 4, and discuss some observations from our case study in Section 5. We conclude the paper by summarizing our contributions in Section 6.

2. RELATED WORK

As mentioned, advances in Natural Language Processing (NLP) has given the possibility of automatic summary grading, such as the successful systems that use Latent Semantic

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Analysis (LSA) [3]. LSA treats each essay as a matrix of word frequencies and applies singular value decomposition (SVD) to the matrix to find an underlying *semantic space*. It then represents each to-be-graded essay in that space as vectors and assesses the cosine similarity between the essay and the graded or standard essays or the text students read. The cosine similarity can be transformed to the grade. LSA-based systems have shown to be very accurate and with significant correlation when compared with human grading.

SAGrader [1] provides a representation of the domain from which the summary is going to be made by combining a pattern matching approach with semantic networks. SAGrader analyzes students' essays and then provides students with a grade, and limited feedback that indicates the items a student mentioned and those they did not based on the domain representation. However, SAGrader does not provide feedback on any additional or inaccurate content a student may have included.

Following the ideas presented by SAGrader, it seems promising that Ontology-based Information Extraction (OBIE) can be used for automatic summary and essay grading because OBIE provides a domain representation to help identify concepts and relationship over free text, which is a similar task to grading summaries and essays. Using OBIE for automatic summary and essay grading offers the same advantages that semantic networks offer, such as the possibility of generating feedback and no need of gold standard summaries. Also, ontologies are more expressive than semantic networks by allowing to represent disjointness and negations, which cannot be done by semantic networks [8].

To accurately evaluate a summary and provide meaningful feedback, we need to identify the correct and incorrect statements that are in it. However, to the best of our knowledge, no research has been done in OBIE to identify and/or extract statements that are logically inconsistent with respect to the ontology. An ontology can become logically inconsistent through changes made to it [4]. Research has been done to prevent inconsistency [2], and to eliminate inconsistency by detecting its source [4, 11]. Although the mentioned work does not analyze text inconsistency based on an ontology, it does give an insight on how this problem could be approached. Since the statements of a summary should be entailed from the domain ontology, if a statement of a summary is incorrect, it will be inconsistent with the ontology. So, understanding how ontology inconsistency is managed can lead to mechanisms to identify and extract incorrect summary statements.

3. METHODOLOGY

In this section, details regarding the design and implementation of the proposed OBIE method are presented.

3.1 Data Collection

The student summaries we will use in this paper were collected in an earlier study that looked at the use of electronic strategies (eStrategies) for reading comprehension for college students [9]. The study included 18 subjects with a range of reading abilities (from high to low). As part of the study, adult students were asked to read four 500-word passages that were drawn from introductory college science textbooks, then provide oral summaries of each article. The oral summaries were manually transcribed into the text form. The study produced a two-part score for each student summary that was based on (a) the LSA score, and (b) a human expert score. Both scores are numeric.

Element type	Number of element
Concepts	49
Relationships	35
Subclass relationships	16

Table 1: Statistical information about the ontology.

Our work starts with 18 summaries of one of the articles, Ecosystems, and the scores for each summary. Our focus is on using our OBIE technique to grade the summaries and compare our results with those of the study (both LSA and human grader).

3.2 Automatic Grading and Error Detection

We have built an OBIE system to do summary grading. The OBIE system follows the component based architecture defined by Wimalasuriya and Dou in [12].

For the present work, the following components have been selected to construct an OBIE automatic grading system:

1. Ontology: Provides formal representation of concepts and relationships of a domain.
2. Preprocessors: Converts text into a format that can be processed by components of the next phase, information extractors.
3. Information Extractors: Performs extractions with respect to a specific class or a property of an ontology (and considered the most important component in the architecture.)

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

3.2.1 Ontology

For the present work, we manually constructed an ontology based on the Ecosystems article. Given that students participating in the study had different levels of background knowledge, the construction was constrained to the explicit facts from the domain article for Ecosystems and does not include facts from the entire domain of ecosystems.

Because of the strict construction criteria and the nature of the Ecosystem article (an introduction to a domain), the ontology is mainly a list of important concepts and relationships (see the statistics in Table 1).

3.2.2 Preprocessing

In order to simplify the complexity of the summaries and to obtain the best performance of the Information Extraction process, a preprocessing stage has been defined. The preprocessing stage considers completing sentences, eliminating non-informative words, and correcting misspellings.

3.2.3 Extractor selection

The two main approaches for extracting information from texts are extraction rules and classification techniques [13]. Extraction rules are based on regular expressions that capture specific types of information. With classification techniques, the method tries to identify if a sentence contains the information sought or not. Extraction rules are simple to design but do not scale well while classification techniques

scale well but require large data sets for training and testing [13]. Since the data size (i.e., number of summaries) for the present work is small, extraction rules are selected as information extractors for the summary grading system.

3.3 Extraction Rules

From the summaries, we identified three types of statements made by the students: correct statements, incorrect statements, incomplete statements.

3.3.1 Rules for correct statements

An ontology formally defines the concepts and relationships in a domain. The relationships can be seen as triples of the form *concept1 relationship concept2*. The triple can be mapped to the typological form of a sentence (*subject verb object*), where *subject* maps to *concept1*, *verb* maps to *relationship* , and *object* maps to *concept2*.

This leads to an extraction rule for each relationship. Since properties of a concept are inherited by its sub-concepts or by its equivalent concepts, we consider the use of first order logic (FOL) rules to combine sets of axioms from the ontology into a smaller set of logical rules to avoid the creation of an oversized set of extraction rules. The resulting logical rules contain concepts and properties from the original set of ontological axioms. In other words, the set of original axioms entail the new logical rules, which is an expanded representation of the original set.

Rules for correct statements can identify which concepts and relationships are presented in the summary, and which are not presented. A total of 31 extraction rules are created from the ontology. Similar to SAGrader [1], the feedback tells how much of a student's summary is contained in the ontology and how much is missing from it.

3.3.2 Rules for incorrect statements

If we consider that statements in a summary should be entailed from the domain ontology, an incorrect statement will be inconsistent with the ontology. Wang et al. [11] proposed a heuristic to identify the cause of inconsistency in an ontology based on common errors they observed in tutorials and workshops regarding ontology creation. They proposed a set of rules to detect these common errors (properties with conflicting domain or range, ignoring disjointness between classes, and conflicting axioms through propagation of axioms) in inconsistent ontologies.

Common errors are having properties with conflicting domain or range, ignoring disjointness between classes, and conflicting axioms through propagation of axioms. Following the common errors identified by Wang et al. [11] and the constraints presented in the ontology, we can create a set of logic rules on inconsistency. The extraction rules created from the inconsistency logic rules should be able to help identify incorrect statements.

From the consistent logic rules, approximately 84 rules on inconsistency can be derived. However, only 16 extraction rules for incorrect statements are used since this small set already covers almost all the incorrect statements made by 18 students.

3.3.3 Rules for incomplete statements

Statements that include a concept or relationship that is not defined in the ontology are considered incomplete. The most frequent type of incomplete statements is related to a relationship between two concepts but one of them is not in the ontology.

To identify incomplete statements, the extraction rules for incompleteness look for statements in the summaries that have an unknown element. The extraction rule that implements the logic rule for incompleteness checks that if in a sentence an element of a stated relationship is not listed in the ontology, then the sentence is incomplete with respect to the ontology.

	Type of Extraction Rules		
Metric	Correct	Incorrect	Incomplete
Precision	91.9%	97.4%	66.67%
Recall	83.3%	88.63%	80%
F1	87.4%	92.8%	72.7%

Table 2: Performance of OBIE

3.4 Grading metrics

The human grading of a summary or essay usually takes into account aspects such as relevance of concepts presented in the summary, if all of main concepts or ideas are presented in the summary, length of the work, and the fluency of the writing. The present work will consider the first two aspects mentioned, plus the amount of relevant information that is present in the summary.

Three metrics are defined to measure the quality of the summaries; each of them is related to one of the three mentioned aspects.

- **Relevance**: This metric considers what part of the summary is related to the article read by the students. The metric provides a ratio of how much of the summary can be matched with the extraction rules.
- **Completeness**: This metric considers how much of the article is contained in the summary. This metric indicates how many rules are matched in the summary.
- **Importance**: This metric gives a weight to each relationship, so that if a summary has most important relationships then it has a better grade than if the summary contains only the less important relationships.

4. EXPERIMENTS

Two aspects must be evaluated when trying to determinate the performance of the proposed system. First, since the core of the grading system is an OBIE system, the performance of the extraction must be evaluated to determine the performance obtained by OBIE. Second, we would like to compare the type of detailed scoring we obtain against that of what the human grader provided. Unfortunately, we only have the human grader's composite score, so we cannot do a detailed analysis of our findings on correctness, incorrectness, and incompleteness against the expert's score. We will take the next best approach, and compare numeric scores. This assumes the expert's scores are based on correctness, incorrectness, and incompleteness.

4.1 OBIE Performance

Table 2 provides the performance of the OBIE system itself which is measured with the metrics Precision, Recall, and F1 measure. While Precision measures how much of the

ID	Sohlberg et al. study		OBIE grading metrics		
	Teacher Grade	LSA	Relevance	Completeness	Importance
STIR10	12	0.969	0.263	0.29	0.315
STIR12	14	0.969	0.6	0.29	0.315
STIR13	8	0.835	0.25	0.096	0.054
STIR15	4	0.931	0	0	0
STIR17	15	0.83	0.368	0.161	0.205
STIR19	17	0.785	0.206	0.161	0.136
STIR2	10	0.759	0.038	0.032	0.013
STIR20	8	0.822	0.25	0.322	0.301
STIR22	12	0.741	0.218	0.258	0.26
STIR23	8	0.818	0.473	0.29	0.315
STIR24	5	0.862	0.111	0.0322	0.054
STIR25	8	0.811	0.375	0.161	0.191
STIR26	3	1.148	0	0	0
STIR27	12	0.811	0.285	0.096	0.109
STIR28	13	0.868	0.578	0.387	0.438
STIR33	7	0.811	0.222	0.064	0.082
STIR4	10	0.734	0.181	0.129	0.109
STIR7	8	0.723	0.193	0.258	0.273

Table 3: Grades of 18 summaries by human grader (Teacher Grade), NLP based grading method (LSA), and OBIE (Relevance, Completeness, Importance).

extraction is correct, Recall measures how complete is the extraction. The F1 measure is the average between Precision and Recall that provides an overall measure of the system.

4.2 Summary grading by OBIE

Using the set of 18 summaries from the Sohlberg et al. study [9], OBIE provided numeric scores for each. The three previously mentioned grading metrics are presented in Table 3 where they are compared to human grading and LSA based grading [9].

Table 4 presents the Spearman correlation between the OBIE grades (Relevance, Completeness, and Importance) and grades from the Sohlberg et al. study (Expert Grade and LSA) [9]. Spearman's correlation measures the dependency between variables described by a monotonic function.

	Expert Grade	LSA
LSA	-0.163	
Relevance	0.531	0.176
Completeness	0.547	-0.018
Importance	0.559	0.016

Table 4: Spearman correlation between grading metrics ($p < .05$)

5. DISCUSSION

5.1 Automatic summary grading

Because the ranges of grades from human graders, LSA, and our systems are totally different, we only conducted correlation studies among them. We found that the grades from our OBIE system resulted in a positive correlation with human grading. In other words, agreement was suggested between scoring by the human grader and OBIE. On the other hand, no correlation occurred between the LSA and OBIE grades, as seen in Table 4. Interestingly, LSA grading and human grading were also not positively correlated. The most straightforward answer is that LSA does not address incorrect statements. Our system does. Given that we found that 75% of the summaries contained at least one error, the divergence from LSA is not surprising.

STIR33 Ecosystems are composed of different types of living organisms. There are herbivores, carnivores, detritivores and omnivores. Detritivores eat inorganic matter or non-living matter. Omnivores eat everything. Herbivores eat meat and other organisms. And herbivores eat vegetation.
STIR26 Carnivores are fish. And I figure out what to say in my head.

Table 5: Example of summaries.

5.2 Error Detection

Examination of two examples help to elucidate the effect of error detection by OBIE. STIR33 has a grade of 7 from the human and .811 from LSA. The OBIE score is [.222, .064, .082]. OBIE found a number of errors in this summary, including:

1. Detritivores do not eat inorganic matter.
2. Omnivores eat only plants and animals. They do not eat organic waste or fragments of dead organisms.
3. Herbivores eat plants.

Note that the original human grader's score was 17 for this summary. Our team contacted the grader to try to gain

insight into her high score given the number of errors we found. Our note to her prompted her to look at her raw scores again and find a typo - her raw score was 7 not 17. We used her change in Table 3.

We can speculate that the high LSA score is because of many article concepts and relationships being mentioned, whether they are rightly or wrongly stated.

For STIR26, LSA provides a high-water-mark in scoring with 1.148. The human score is 3 (low-water mark). The OBIE score is 0 across the board (low-water mark). OBIE found 2 errors in the first statement:

1. Carnivores are not a sub-class of fish.

2. Assuming a transpose error, the statement "Fish are carnivores." is also incorrect. In essence, one cannot find a way to patch a relationship between fish and carnivores without creating new concepts.

The second statement is marked as non-relevant by OBIE - it has no concepts or relationships that occur in the ontology. We can only guess why LSA gave this summary its highest score.

A key point here is that we are comparing numeric scores to make sure we are in the right ballpark with the human grader - lacking her score breakout, we have nothing other to compare against. However, the key goal of our work is to provide not a numeric score to a student, but feedback. And OBIE provides the raw basis for such feedback. The errors that we delineate are the grist for helping a student rework his or her conceptualization. While we did not attempt to provide tutoring as part of this study, it is one of our overall goals. Beyond simply showing errors, we can foresee a more sophisticated approach to hold tutoring dialogs similar to that reported in [10]. Recognizing conceptually correct, incorrect, and incomplete ideas at the statement level is what drives our work.

6. CONCLUSIONS

We have presented a novel Ontology-based Information Extraction system for grading summaries that correlates well, on a numeric level, with an expert human grader. At the same time, our approach provides meaningful feedback to the students about the incorrect statements they have made in the summaries.

In terms of the underlying OBIE system, there are several goals we have that are discussed below. We consider integrating text taxonomy as a complement to the ontology to allow a better link between the domain knowledge and importance of each idea in the text. Bring some automation to the process would widen the available material that we could support with our approach, such as machine learning (e.g., classification) techniques as information extractors, or automatically generating extraction rules.

7. ACKNOWLEDGMENTS

We thank Daya C. Wimalasuriya for discussions on the component based OBIE systems. 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.

8. REFERENCES

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

[2] G. Flouris, Z. Huang, J. Z. Pan, D. Plexousakis, and H. Wache. Inconsistencies, negations and changes in ontologies. In *AAAI*, pages 1295–1300, 2006.

[3] P. W. Foltz, D. Laham, and T. K. Landauer. Automated essay scoring: Applications to educational technology. In *EdMedia*, pages 939–944, 1999.

[4] P. Haase and L. Stojanovic. Consistent evolution of owl ontologies. In *ESWC*, pages 182–197, 2005.

[5] J. D. Karpicke and H. L. Roediger. The Critical Importance of Retrieval for Learning. *Science*, 319(5865):966–968, 2008.

[6] T. K. Landauer, D. Laham, and P. W. Foltz. Learning human-like knowledge by singular value decomposition: a progress report. In *NIPS*, pages 45–51, 1998.

[7] M. A. McDaniel, H. L. Roediger, and K. B. McDermott. Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, 14(2):200, 2007.

[8] S. J. Russell and P. Norvig. *Artificial intelligence : a modern approach.*, chapter Knowledge representation, page 454. Prentice Hall, 3rd edition, 2010.

[9] M. Sohlberg, G. Griffiths, and S. Fickas. The effect of electronically delivered strategies on reading after mild-moderate acquired brain injury. (Submitted to) *American Journal of Speech-Language Pathology*, 2012.

[10] K. VanLehn, A. Graesser, G. Jackson, P. Jordan, A. Olney, and C. Rosé. When are tutorial dialogues more effective than reading? *Cognitive Science*, 31:3–62, 2007.

[11] H. Wang, M. Horridge, A. Rector, N. Drummond, and J. Seidenberg. Debugging owl-dl ontologies: A heuristic approach. In *ISWC*, pages 745–757, 2005.

[12] D. C. Wimalasuriya and D. Dou. Components for Information Extraction: Ontology-Based Information Extractors and Generic Platforms. In *CIKM*, pages 9–18, 2010.

[13] D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. *J. Inf. Sci.*, 36:306–323, 2010.