Spring 2014

A Computer-Based Approach For Identifying Student Conceptual Change

Junchao Yan
Purdue University

Graduate School ETD Form 9
(Revised 01/14)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Junchao Yan

Entitled
A COMPUTER-BASED APPROACH FOR IDENTIFYING STUDENT CONCEPTUAL CHANGE

For the degree of Master of Science

Is approved by the final examining committee:

Dr. Alejandra J. Magana

Dr. Bedrich Benes

Dr. Grant P. Richards

To the best of my knowledge and as understood by the student in the *Thesis/Dissertation Agreement. Publication Delay, and Certification/Disclaimer (Graduate School Form 32)*, this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Dr. Alejandra J. Magana

Approved by Major Professor(s): _____

Approved by: Jeffrey L Whitten 04/29/2014

Head of the Department Graduate Program Date

A COMPUTER-BASED APPROACH FOR IDENTIFYING STUDENT CONCEPTUAL CHANGE

A Thesis

Submitted to the Faculty

of

Purdue University

by

Junchao Yan

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

May 2014

Purdue University

West Lafayette, Indiana

Dedicated to my wife and my family.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude to my advisor, Dr. Alejandra J. Magana, who continuously provides tremendous support and guidance during my graduate studies and research. Thank you for giving me the opportunity to study here and leading me into this area. I also would like to thank my committee members, Dr. Benes and Dr. Richards, for their insightful advice and continuing to encourage me throughout my studies. I thank everyone in our group, especially Camilo Vieira Mejia, Karla Sanchez, and Miguel Angel Ruiz. Thanks for your friendship and for your generous help and support during my time here.

I would like to extend my gratitude to my friends: Luojie Xiang, Shunrang Cao, Na An, Pengcheng Xue, Hui Yan, Anyi Li, Qian Zhang, Hang Lu, Liting Xu. Thank you all for making the Midwestern life so colourful.

I owe my deepest gratitude to my wife, Xi Chen. Thank you for your love and always being there for me. You have made me the happiest I have ever been. Lastly, I am deeply grateful to my father Yimin Yan and my mother Ying Liu. Thank you for your unconditional love and support throughout my life. I would not have been able to accomplish anything without every one of you. Thank you for everything.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

ABBREVIATIONS

MOOC	Massive Online Learning Course
TTCI	Thermal and Transport Concept Inventory
NLP	Natural Language Processing
LSA	Latent Semantic Analysis
SVD	Singular Value Decomposition
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
POS	Part-of-Speech

GLOSSARY

Conceptual change	a shift across ontological categories (Chi, 2005)
Emergent process	a process with a random and simultaneous pattern (Chi & Roscoe, 2002)
Sequential process	a process with a causal and dependent pattern (Chi & Roscoe, 2002)

ABSTRACT

Yan, Junchao M.S., Purdue University, May 2014. A Computer-based Approach for Identifying Student Conceptual Change. Major Professor: Alejandra J. Magana.

Misconceptions are commonly encountered in many areas of science and engineering where a to-be-learned concept conflicts with prior knowledge. Conceptual change is an approach for identifying and repairing the misconceptions. One of the ways to promote student conceptual change is providing students with ontological schema training. However, assessment of conceptual change relies on qualitative analysis of student responses. With the exponential growth of qualitative data in the form of graphical representations or written responses, the process of data analysis relying on human experts has become time-consuming and costly. This study took the advantages of natural language processing and machine learning techniques to analyze the responses effectively. In addition, we identified how students described complex phenomena in thermal and transport science and compared the differences of descriptions between students who took certain training courses to address misconceptions by means of ontological schema training and those who were exposed to a different course about the nature of science. After comparing the effectiveness of three different text classification methods - query-based approach, Naïve Bayes classifier, and support vector machine (SVM) for identifying conceptual change, SVM classifier was chosen to assess student responses from a corpus collected by Streveler and her research group in previous studies (Miller et al., 2011). Based on the automatic assessment for student conceptual change, this research found that training students with appropriate ontological schema would promote the conceptual change.

CHAPTER 1. INTRODUCTION

Technology has offered innovations to help improve education in multiple ways. For instance, educational technologies are not only making learning more accessible to everyone who wants to learn by providing high-quality massive online open courses (MOOC), but also advancing techniques in data analysis to help educators better understand the process of learning. To keep abreast of the fast changes, the Department of Education of United States published a report in 2012 named "Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief" (Bienkowski, Feng, & Means, 2012), discussing the potential efforts and challenges of applying data mining and data analytics in education. With the deluge of learning data, educators and educational researchers should be prepared to identify new forms of computer-based assessment that can help them better identify how students learn, what students know, and furthermore, what they do not know. One of the educational research areas that is mature enough to take advantage of these new technologies is the body of *knowledge and research in the area of conceptual change*.

Conceptual change is an approach for identifying and repairing misconceptions in science and engineering (Miller et al., 2011). Misconceptions are commonly encountered situations in many areas of science and engineering where a to-be-learned concept conflicts with prior knowledge. Many studies attempted to understand and explain the mechanism of conceptual change (Carey, 1985; Chi, Slotta, & De Leeuw, 1994; Pintrich, Marx, & Boyle, 1993; Posner, Strike, Hewson, & Gertzog, 1982; Vosniadou & Brewer, 1992).

One of the ways educational researchers can identify misconceptions is by means of concept inventories (Miller, Streveler, Nelson, Geist, & Olds, 2005). Concept inventories are a set of questions designed to evaluate student knowledge of

a group of certain concepts (Hestenes, Wells, & Swackhamer, 1992). Another means is by qualitatively analyzing student explanations and rationale for explaining scientific phenomena by open-ended questions (Miller et al., 2011). All these data have become important to identify what students know and, in a way, how they learn. Computer-based assessment techniques have started to be used to analyze student responses to concept inventories. For instance, ciHUB.org is an online resource funded by the National Science Foundation (NSF), providing state-of-the-art assessment tools that allow faculty to identify patterns of conceptual understanding and misconceptions by means of quantitative data (ciHUB, 2011). However, with the exponential growth of qualitative data in the form of graphical representations or written responses, the process of data analysis relying on human expert has become time-consuming and costly. Therefore, automated methods for analyzing and assessing student responses are needed.

Learning analytics (Siemens & Baker, 2012), which is an emerging and fast growing interdisciplinary field, combines education, natural language processing, and data mining to address this problem. Learning analytics is defined by Society of Learning Analytics Research (SoLAR) as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (*Society for Learning Analytics Research*, 2011). Being driven by the rise of big data and MOOCs, learning analytics provides opportunities to analyze qualitative data effectively, and to help researchers better understand student learning processes. In addition, natural language processing has been extensively studied in automatic scoring of essays and summaries (Doddington, 2002; He, Hui, & Quan, 2009; Lin & Hovy, 2003; Noorbehbahani & Kardan, 2011), educational text classification (Petersen & Ostendorf, 2009), based on shallow features. However, when assessing student short responses, deeper semantic features are needed (Leacock & Chodorow, 2003; Mohler, Bunescu, & Mihalcea, 2011; Nielsen, Ward, & Martin, 2008; Pulman & Sukkarieh, 2005).

1.1 Scope

The scope of this research is to identify how students describe complex phenomena in thermal and transport science and to compare the differences of descriptions between students who took certain training courses to address misconceptions by means of ontological schema training and those who were exposed to a different course about the nature of science. In addition, this research compared the effectiveness of three different text classification methods in identifying conceptual change. One of the methods was chosen to assess student responses from a corpus collected by Streveler and her research group in previous studies. The corpus contains student answers to multiple-choice questions, which were coupled with open-ended questions that prompted students to explain their rationale for their responses, from the Thermal and Transport Concept Inventory (TTCI) (Miller et al., 2011; Yang et al., 2012).

1.2 Significance

Conceptual change is a crucial process of learning (Chi et al., 1994). Extensive research has been done to understand how to promote conceptual change and repair misconceptions (Miller et al., 2011; Slotta & Chi, 2006; Yang et al., 2012). To better understand this process, a large experiment has been conducted using qualitative and quantitative methods (Yang et al., 2012). While the analysis of quantitative data was a straight forward process, the analysis of qualitative data may be time-consuming. Previous work has applied natural language processing (NLP) and data mining techniques to assess student essays and predict student performance. Nevertheless, none has been able to explore the use of NLP in assessing student conceptual understanding based on short responses. By applying state-of-art NLP techniques, this research gave a better understanding of student level of conceptual change in thermal and transport science, and how students repair misconceptions by building a new mental representation or schema.

1.3 Research Questions

- How do students describe complex phenomena with and without the ontology schema training?
- What is the effectiveness of using computer-based approach for identifying student conceptual change?
- Do students demonstrate conceptual change after being exposed the ontological schema training based on computer-based assessment?

1.4 Assumptions

The assumptions for this study include:

- Students had basic knowledge of the concepts from thermal and transport science.
- Students were honest when answering questions.
- Students did understand the learning materials that were provided in this study.
- There were consistent patterns in student answers.

1.5 Limitations

The limitations for this study include:

- The participants of this study were only from one engineering university.
- This study was limited to the number of participants.
- The training set used in this study was small.

- This study only investigated the conceptual change of students in thermal and transport science.

1.6 Delimitations

The delimitations for this study include:

- Data collection was not performed as part of this study.
- Mental representations other than emergent process and sequential process were not considered in this study.
- Concepts other than diffusion, heat transfer, and microfluidics in thermal and transport science were not considered in this study.

1.7 Summary

This chapter provided scope, significance, research questions of this study. It also covered the assumptions, limitations, and delimitations. The next chapter provides a view of literature including conceptual change and learning analytics.

CHAPTER 2. LITERATURE REVIEW

2.1 Conceptual Change

Conceptual change is a process of changing existing conceptions. Early research has been done to understand and explain the nature of conceptual change (Carey, 1985; Chi et al., 1994; Pintrich et al., 1993; Posner et al., 1982; Strike & Posner, 1992; Vosniadou & Brewer, 1992). To answer the basic question of how student conceptions changed when they were trying to learn new ideas and concepts, Posner et al. (1982) proposed a general model of conceptual change based on the philosophy of science, where they referred accommodation as the situation when student current knowledge is not able to explain new phenomena. They suggested that accommodation was likely to happen when the four conditions were met including (1) dissatisfaction with existing concepts, (2) the new concept should be explainable, (3) the new concept should be intelligible, and (4) the new concept should be able to be extended. While Posner et al. (1982) explained conceptual change in a conceptual ecology perspective, others had described it in terms of mental models (Vosniadou & Brewer, 1992), and ontological shifts (Chi et al., 1994).

Vosniadou and Brewer (1992) used earth's shape as an example to investigate children conceptual change at the mental model level. They found that the initial mental model slowly changed to a mental model of a sphere earth, which meant that the instructions led the children to form a synthetic model that adjusted the new information to their initial mental model.

One of the most influential theories of conceptual change was proposed by Chi et al. (1994), who claimed that concepts are difficult to learn due to the mismatch between an ontological status and a naive conception. A follow-up

experiment conducted by Slotta and Chi (2006) revealed that providing training materials about the target ontology before the formal instructions could help novices develop a better understanding of complex phenomena.

Streveler, Litzinger, Miller, and Steif (2008) discussed the significance of conceptual knowledge in engineering sciences and provided examples to illustrate the conceptual difficulties in mechanics, thermal science, and direct current electricity. They demonstrated two basic types of difficulties (1) understanding of basic quantities, (2) understanding of relationships among the basic quantities. The authors also pointed out that in higher education a limited research on conceptual knowledge and misconception has been conducted. Therefore, a lot of questions remain unanswered, for instance, why students have difficulty with elementary quantities and basic relationships? How does conceptual knowledge evolve as a learner moves from novice toward expert performance? What are the factors that affect conceptual change? In a follow-up experiment, Yang et al. (2012) designed computer-based online learning modules to help students develop new mental representations to promote conceptual change based on Chi's theory of ontological schema training. The modules were validated through a series of experimental designs. In their study, the participants were 60 undergraduate students, and were randomly assigned to two groups, of which the experimental group was given instruction based on the ontological schema training approach (Slotta & Chi, 2006), and the control group was exposed to instruction related to the nature of science. The results from the TTCI showed that students in the experimental group understood the concepts from diffusion and microfluidics better compared with students from the control group. However, the reasons why no significant differences were found between two groups in topic of heat transfer need to be investigated. Therefore, there are needs for further data analysis.

2.1.1 Ontological Schema Training

Ontological schema training is an approach that promotes student conceptual change by providing appropriate schema related the to-be-learnt concepts(Slotta & Chi, 2006). Previous research has found scientific concepts are generally related to two distinct categories of processes, *sequential processes* and *emergent processes* Chi (2005). Emergent processes are the situations when elements of a system interact in a random and simultaneous pattern, where direct processes happen in a causal and dependent pattern. However, misconception may happen when students falsely recognize emergent processes as sequential processes (Chi, 2005). Therefore, in order to help student learn the emergent processes, rich examples with the properties of emergent processes need to be provided. This learning process is referred as ontological schema training, which has been successfully used to help students repair their misconception (Chi, Roscoe, Slotta, Roy, & Chase, 2012; Miller et al., 2011; Slotta & Chi, 2006; Yang et al., 2012).

2.2 Learning Analytics

Educational data mining and learning analytics are the areas that comprise techniques from machine learning, data mining, and natural language processing to extract useful information from large educational datasets. One key difference of these two areas is that educational data mining focuses on automatic discovery while learning analytics aims at leveraging human judgement through data visualization and data analysis methods (Siemens & Baker, 2012). Compared with the emerging areas - learning analytics and educational data mining, computer-based assessment has been developed for years, and been proved to be an effective tool for learning (Thelwall, 2000). Leaning on modern computer technology makes it possible to assess the performance of large groups of students with effective methods. Automatic assessment methods can be grouped into three categories: latent semantic analysis (LSA) based (Landauer, Foltz, & Laham, 1998; Landauer,

Laham, Rehder, & Schreiner, 1997), N-gram based (Lin & Hovy, 2003; Noorbehbahani & Kardan, 2011; Pérez, Alfonseca, & Rodríguez, 2004) and integrated methods (He et al., 2009; Pérez et al., 2005), among which LSA has been shown to be a powerful tool in cognitive science.

LSA is not only a theory but also a method for discovering and representing the meaning of words. It was first introduced for indexing documents in information retrieval in the late 1980's (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990; Landauer & Dutnais, 1997; Landauer et al., 1998). LSA reveals the correlation between terms and documents by decomposing the term-document matrix with singular value decomposition (SVD). After choosing the most related terms, the cosine similarity between reference and student answer vectors can be calculated.

Olde and Franceschetti (2002) applied LSA to analyze student responses for physics questions. Participants in this study were 120 students, and each of them answered 10 questions, where all the answers were evaluated by four experts with a five point scale. To compare the five different physics corpora which was used to train the LSA space, correlation between LSA scores and expert scores was computed. Based on the result, LSA provided reasonable scores that were similar with the experts'. In addition, no difference was found on the correlation between LSA and human judgments scores whether the irrelevant information was eliminated from the corpora or not.

Bilingual evaluation understudy (BLEU) is a modified n-gram methods for automatic evaluation method of machine translation (Papineni, Roukos, Ward, & Zhu, 2002). More importantly, a strong correlation between these automatically generated scores and human judgements of translation quality has been reported. Inspired by this method, Pérez et al. (2004) applied it in assessing short essays written by students, namely Evaluation Response with BLEU (ERB). In this case, the student answers were treated as the candidate translations that were supposed to be scored, and the reference translations were the answers given by the instructors. Therefore, the closer a student's answer was to the instructor's, the

better it was. In the result, it was shown that the modified algorithm could perform a reasonable correlation with the human assessments.

Similar to BLEU, n-gram co-occurrence was first introduced by Doddington (2002) to evaluate the quality of machine translation. Lin and Hovy (2003) conducted a study on comparing n-gram co-occurrence and BLEU for automatic summary evaluation. As a result, unigram co-occurrence statistic was proved to be an accurate automatic evaluation metric as it achieved high correlation with human judgements. Noorbehbahani and Kardan (2011) modified the BLEU algorithm to make it suitable for the assessment of free text answers. The author pointed out some drawbacks of BLEU including (1) exact match, (2) equally weighted, (3) multiple reference, and (4) simple penalty. Therefore, a refined evaluation score was proposed using equation shown in 2.1.

$$M\text{–}BLEU = exp[\sum_{i=1}^{N} w_n \log(WP_{ra}(n))] \tag{2.1}$$

The notation WP_{ra} is the weighted n-gram precision. In this method, the most similar reference answer was chosen first. Then it calculated the similarity score based on measurements of M-BLEU, BP and common-words order similarity between a selected reference and a student answer. The method was showed to have a 85 percent correlation with average expert scores.

To perform both syntactic and semantic analysis, Pérez et al. (2005) incorporated ERB with LSA. In this method, ERB and LSA scores were calculated by equation 2.2 and equation 2.3, respectively.

$$ERB(a) = MBP(a) \times exp(\sum_{n=0}^{N} \frac{log(MUP(n))}{N}) \tag{2.2}$$

$$LSA(a) = \frac{\sum_{\mathbf{r}_i \in R} \cos(\mathbf{r}_i, \mathbf{a})}{2|R|} + 0.5 \tag{2.3}$$

The notation a, N, $\mathbf{r}_i$, $\mathbf{a}$ denote the evaluated answer, the length of n-grams, the reference vector and the answer vector, respectively. Then, a combined score can be calculated as follows:

$$COMB(a) = \alpha ERB(a) + (1 - \alpha) LSA(a) \tag{2.4}$$

α is a parameter used to adjust the weight. The result showed that the mean correlation to the human judgements had reached 50 percent.

He et al. (2009) combined the LSA and n-gram co-occurrence methods to improve the accuracy in automatic summary assessment. In this case, LSA was used to find the "most correlated terms" from both students" answers and reference answers. After calculating the cosine distance between them, the n-gram co-occurrence was applied to match the summary. The final score was calculated as:

$$\frac{LSA\ score + N\text{–}gram/co\text{–}occurrencescore}{2} \tag{2.5}$$

As a result, this method achieved 96 percent accuracy compared to 87 percent for BLEU algorithm.

2.3 Summary

This section examined the theory of conceptual change, the power of latent semantic analysis and its applications in education research. It has demonstrated that CAA methods are able to achieve reasonable accuracy and correlation to human judgements. In addition, text categorization methods provide an effective way when we need to group or label the students' responses during the assessment. However, it also shows that there are still remained gaps for computer-based assessment of conceptual understanding.

CHAPTER 3. THEORETICAL FRAMEWORK

As conceptual change is an interaction between student prior knowledge and what they are about to learn, a misconception occurs when their prior knowledge conflicts with what they are learning. Misconceptions have been reported in various areas in engineering and science education, of which some are very robust and hard to repair. An explanation is that the to-be-learned concept is ontologically assigned to a false category (Chi, 2005).

With the hierarchical structure of conceptual knowledge, an ontological category can be defined by shared ontological attributes of similar concepts. Therefore, ontological distinction happens in a situation where two categories do not have many common ontological attributes. In the context of conceptual change in science and engineering, Chi et al. (1994) identified three ontological trees - *matter* (things), *processes*, and *mental states*, in which a hierarchy of subcategories are embedded. They found that many scientific concepts belonged to an ontological category that could be referred to as constraint-based interaction, a subcategory of *processes*. Some attributes of this category could be summarized as: no beginning or end, no progression, causal, uniform in magnitude, simultaneous, static, on-going, and so on. It is also worth pointing out that constraint-based interactions share some components from the ontological category of matter. This may explain why there are so many misconceptions in engineering and science education. Chi et al. (1994) also suggested that the reason why some misconceptions were robust relates to the difficulty in shifting between two different ontological categories.

As suggested by Slotta et al. (1995), the reasons of why previous research on science instruction has not been successful were lacking of solid theories of conceptual change, as well as the difficulty of assessing conceptual change. In their study, they qualitatively analyzed verbal explanations of physics phenomena (light,

heat, electrical current) by both novices and experts. They found that there was a different pattern of predicate used by novices and experts, which reflected ontological attributes. Therefore, verbal predicates could be used to identify conceptual change.

In a follow-up study, Chi (2005) used concepts of circulation and diffusion as examples to investigate student understanding about complex phenomena. They found that student misconceptions for circulation were non-robust because their direct processes and the correct conception are at the same ontological type. However, when students falsely recognized the emergent processes as direct processes, the misconceptions were robust. Emergent processes happen when a system of constituent elements interact in a random and simultaneous pattern, where direct processes happen in a causal and dependent pattern. According to their findings, Chi (2005) claimed that two additional learning processes - building and ontological shifting, should be required in the instructions.

Following the same procedures, Slotta and Chi (2006) designed experiments to test the hypothesis that providing training materials about the target ontology before formal instruction could help novices develop a better understanding of complex phenomena. The participants in the experiment were 24 undergraduate students, and were randomly assigned to two groups. The difference between the two groups was that the experimental group was given an ontology training module before the instructions on electricity while the control group received training in related science concepts. A comparison of the pre- and post-test scores revealed that students who had the ontology training performed better than the ones who did not. In addition, students who understood training materials better achieved greater gains during the study. The researchers also used as a measure the verbal predicates in student explanations to analyze the ontological associations. However, it is worth pointing out that the use of multiple choice questions in pre- and post-test provided limited understandings of conceptual change.

3.1 Implications For This Study

Slotta et al. (1995) proposed a methodology for evaluating the use of predicates as a measure for conceptual change. This same methodology was also used in following research studies of conceptual change (Chi, 2005; Slotta & Chi, 2006; Yang et al., 2012), which suggests that there is a strong pattern in the use of verbal predicates. An example of the use of predicates for substance and process is shown in Table 3.1.

Meanwhile, classification, which is an example of pattern recognition in machine learning, aims to assign an item to one of the predefined categories. Based on previous findings, it is strongly suggested that assessment based on recognizing the pattern of the use of verb predicate from student responses can be viewed as a problem of text categorization. Therefore, the implications of the theoretical framework for this study is that we will use text categorization methodologies to analyze student verbal predicates as a way to identify if their verbal explanations are more substance oriented or process oriented. We expect that by identifying and categorizing student verbal explanations we may be able to identify conceptual change.

Table 3.1
Example of taxonomy of predicates (Slotta et al., 1995)

Predicate		Examples
Substance		
	Block	"keeps", "bounces off", "hits", "stops"
	Contain	"holds in", "stores", "keeps in"
	Move	"goes", "leaves", "comes", "flows through"
	Rest	"stops", "stays", "sits"
	Consume	"gets used up", "gets burned up", "burns out", "drains"
	Absorb	"aborbs", "soaks up", "takes in"
	Quantify	"some", "all", "most", "less", "none of", "lots", "little bit", "as much"
	Accumulate	"fills up", "builds up", "adds on", "keeps building"
	Supply	"gives off", "provides", "comes from", "comes out of"
	Equivalent amounts	"the same amount to all of the bulbs", "divides up equally"
Process		
	Transfer	"charged particle moving in an electric field", "the light is a traveling electromagnetic wave"
	Excitation	"energy propagates through", "transfer from one to another"
	Interaction	"a lot of phonon nodes to excite", "need a lot of energy to excite them"
	Equilibrium	"the system finds its way into equilibrium"
	Simultaneous	"they all see at the exact same time"

CHAPTER 4. METHODOLOGY

4.1 Research Framework

The objectives of this study were (1) to find how students describe complex phenomena with and without being exposed to learning materials about thermal and transport science, (2) to develop automatic methods for analyzing student conceptual understanding in thermal and transport science, and (3) using automatic methods to identify whether students can develop a novel mental representation after being exposed to the designed instructions. To achieve the purposes of this study, qualitative data were transformed into numbers to be analyzed. Therefore, this study fell into the category of quantitative research.

4.2 Participants and Procedures

The experiment was repeated three times between 2009 and 2011. A total of 195 students participated in the three sessions of the experiment, with 66 participants in session one, 62 in session two, and 67 in session three. The participants were junior or senior undergraduate students in mechanical, chemical, or materials engineering, who had completed at least one course in heat transfer. They were randomly assigned to two treatment groups, control and experimental. A pre-test was given to these students to determine their prior knowledge about basic heat transfer concepts. Then they were required to complete other three online learning modules, which were about diffusion, heat transfer, and microfluidics.

Different from the control group, the experimental group had the ontological schema training before instructions. The participants were given a training module,

which described the two different kinds of flow processes - emergent and sequential process. In addition, it explained diffusion as an example of emergent process. Meanwhile, the control group was given a same length training module, which described the nature of science without any information about emergent process. The following learning modules for diffusion, heat transfer, and microfluidics were identical for both groups. During the study of each module, students were required to answer some open-ended questions. At last, a post-test was given after all the modules were completed (Yang et al., 2012).

4.3 Dataset

For each experimental session, participants were given tests on three different topics, including "Diffusion", "Heat transfer", and "Microfluidics". Student conceptual understanding was assessed by multiple choice and open-ended questions chosen from TTCI (Miller et al., 2011). In addition, the researchers used student answers to the open-ended questions as the dataset, which consisted of 7,372 responses answered by the students, of which 670 responses were annotated (qualitatively analyzed) by three educational experts.

4.4 Ethical Conduct of Research

This study was approved by the Institutional Review Board (IRB) for conducting research on human subjects. As stated in the application, the user login ID was connected to the response in each single observation, and no other identifiable data was used as part of this study. To maintain the confidentiality of the subjects, the user login ID of every user was replaced by an internal identifier. Furthermore, findings of this study was reported in aggregated form where no user login ID was connected to the data.

4.5 Data Analysis

The textual data was represented by the bag-of-words model, and then was transformed to a co-occurrence matrix, which is also known as term-document matrix. Various term weighting schemas, such as term frequency-inverse document frequency (tf-idf), logarithm tf-idf, and Okapi BM25, were considered (Salton & Buckley, 1988). To identify student conceptual change, automatic text categorization methods including Naive Bayes and support vector machine (SVM) were be applied. Figure 4.1 shows the process of text categorization. Moreover, precision, recall, and F-measure were calculated for the comparison of performance on test data. To evaluate the relationship between the automatic methods and human judgement, results were analyzed using Cohen's kappa, which is a statistical measure for inter-rater agreement. In addition, SVM classifier was applied to investigate whether students demonstrated conceptual change based on their responses. A t-test was performed on the number of responses using emergent language of each student between control and experimental group.

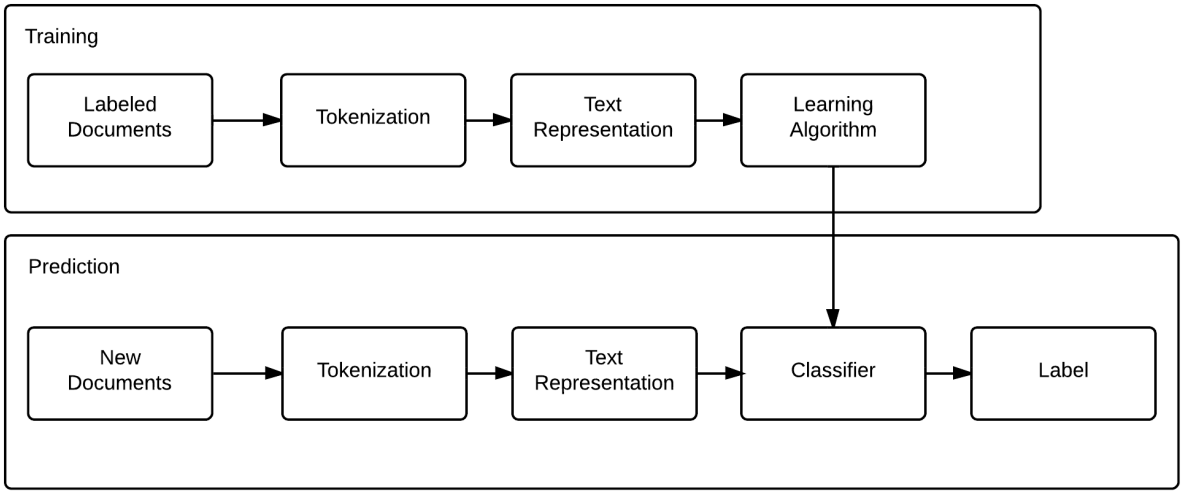

Figure 4.1. Process of Text Classification

4.5.1 Tokenization

Tokenization is a process that breaks a sequence of characters into tokens, such as terms and words (Manning, Raghavan, & Schütze, 2008). It is a widely used approach for pre-processing in NLP. For English, whitespace characters are used to separate tokens. For example, a response, "Diffusion is spreading and mixing of gases or liquids from the random motion of molecules", was tokenized as a set of words ("Diffusion", "is", "spreading", "and", "mixing", "of", "gases", "or", "liquids", "from", "the", "random", "motion", "of", "molecules").

4.5.2 Term Frequency-Inverse Document Frequency

Term frequency-inverse document frequency (tf-idf) is a statistical measure of importance for each term in the corpus, which is also known as term weight in information retrieval (Manning et al., 2008). This weight combines two measures, namely term frequency and document frequency. Term frequency is simply the number of occurrences of a term in a document from the collection while document frequency means the number of documents that contains this term. Term frequency and inverse document frequency are defined as follows:

$$tf_{t,d} = \frac{N_{t,d}}{N_d} \tag{4.1}$$

$$idf_t = \log \frac{N}{df_t} \tag{4.2}$$

The notations t and d represent a term and the document that contains it, respectively. $N_{t,d}$ is the number of occurrence for term t in document d, N_d is the total number of terms in document d, N is the total number of documents in the collection, and df_t is number of documents that contains term t. Therefore, tf-idf is calculated as:

$$tf\text{–}idf_{t,d} = tf_{t,d} \times idf_t \tag{4.3}$$

The value of tf-idf increases as the occurrence of a term in the document increases, but it decreases if this term appears in many documents. After transforming the

term into tf-idf measures, each document in the collection can be represented as a vector.

4.5.3 Classification Approaches

This section briefly introduces the main classification approaches used in this study: Naïve Bayes and Support Vector Machine (SVM).

Naïve Bayes classifier is a probabilistic classifier based on Bayesian theorem. It is a generative model, and it greatly simplifies learning by assuming that features are independent given classes. Although independence is generally a poor assumption, in practice Naïve Bayes often competes well with more sophisticated classifiers. During the model training process, maximum likelihood estimator and Laplace smoothing are applied to find model parameters for a category c, which are calculated as follows:

$$p(t|\theta_c) = \frac{1 + \sum_i^{|c|} tf_{c_i}(t)}{|W| + \sum_i^{|c|} |d_{c_i}|} \tag{4.4}$$

where c represents the category and W is the size of the vocabulary. And the prior probability of category c is:

$$p(c) = \frac{|c|}{\sum_{i=1}^{m} |c|_i} \tag{4.5}$$

Therefore, the prediction of a new input document is:

$$\arg\max_c p(c) \prod_{i=1}^{N} p(tf_i|c) \tag{4.6}$$

Support Vector Machine is a learning algorithm used for classification in machine learning. Each document can be represented as a point in the vector space, where each of its coordinate axes is mapped to a term. To classify the documents, SVM builds a hyperplane to separate the data points from different groups.

4.6 Evaluation

To evaluate the effectiveness, typical metrics such as precision, recall, F_1 score were used. In information retrieval, precision is the fraction of retrieved

documents that are relevant, while recall is the fraction of relevant documents that are retrieved. In statistics, if the null hypothesis is that all and only the relevant documents are retrieved, the precision and recall correspond to fractions of absence of type I and type II errors, respectively. F_1 score is a measure of a test accuracy, which can be calculated as:

$$F_1 = 2\frac{precision \times recall}{precision + recall} \tag{4.7}$$

Cross validation is a model evaluation technique, which can be used for model selection and performance estimation. The simplest cross validation is holdout method, where it separates the labelled dataset into a training set and testing set. Then the training data are used to build the model while testing data are used to predict the output. By comparing the predicted output and correct label by human expert, prediction error can be calculated. For k-fold cross validation, it divides the labelled dataset into k subsets, where one of them is testing dataset and the other k-1 subsets are training set. Then holdout method is repeated for k times, and an average error can be calculated.

4.7 Summary

This chapter provided research framework and methodology used in this study. Next chapter presents the results of this research.

CHAPTER 5. RESULTS

A total of 195 students participated in the three sessions of the experiment, with 66 participants in session one, 62 in session two, and 67 in session three. For each experimental session, participants were given tests on 3 different topics, including "Diffusion", "Heat transfer", and "Microfluidics", with 16, 11, and 5 questions in each topic test, respectively. After removing all answers of the multiple choice questions from the corpus, there were a total of 4921 responses to the open-ended questions for all 3 sessions. Among all responses, 1910 were in the topic of "Diffusion", 2079 were in "Heat transfer", and 932 were in "Microfluidics".

Descriptive statistics of the corpus are shown in Table 5.1. The size of vocabulary for each topic was 1660, 1829, and 1017 unique words, respectively, excluding stopwords, such as "be", "is", "are", and "have". The average length of all responses were 22.50, 16.00, and 19.67 words.

Zipf's law was used to investigate the word distribution across the corpus. The log-log graph of Zipf's law for the corpus is shown in Figure 5.1. The distribution of words for each topic followed the Zipf's law, indicating that a small number of words occurred frequently, while most of the words occurred rarely.

Table 5.1
Descriptive statistics of the corpus

Topic	# of responses	Vocabulary	Average length	Standard deviation
Diffusion	1910	1660	22.50	19.83
Heat transfer	2079	1892	16.00	10.42
Microfluidics	932	1017	19.67	16.58

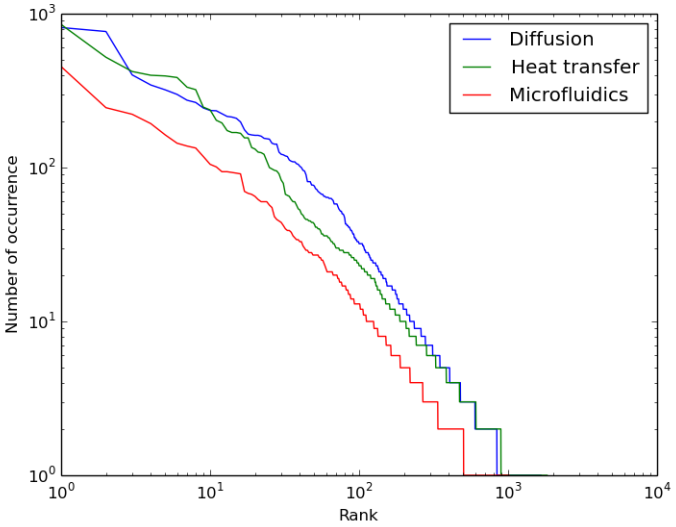

Figure 5.1. Log-log graph of Zipf's law for the corpus

5.1 Use of language with and without ontological schema training

To understand how students described complex phenomena with and without the ontological schema training, responses to each topic were examined using natural language processing (NLP) approaches including tokenization, stemming, and Part-of-Speech (POS) tagging. The most frequent used terms, adjectives, and verbs were identified for each topic. In addition, comparisons were made between the answers from control and experimental group.

5.1.1 Terms

The top 10 words used in student responses are shown in Table 5.2. Results showed that these words were highly correlated with their respective topic. For example, "molecules" and "concentration" are commonly used in Diffusion, same as "heat" and "energy" in Heat transfer, and "dye" and "diffuse" in Microfluidics. All three topics share certain top words, including "water" and "molecules". Nevertheless, the top words for the 3 topics were highly uncorrelated between topics.

Table 5.2
Top 10 words of responses in Diffusion, Heat transfer, and Microfluidics

Diffusion		Heat transfer		Microfluidics	
Words	Occurrence	Words	Occurrence	Words	Occurrence
molecules	921	energy	909	dye	470
concentration	813	heat	851	water	453
water	766	caloric	521	diffuse	245
salt	400	water	421	flow	222
diffusion	344	hot	398	diffusion	193
move	320	transfer	394	molecules	163
dye	299	ice	385	bacteria	144
air	274	drink	333	particles	138
beaker	265	temperature	321	mixing	134
oxygen	244	cube	247	stream	118

5.1.2 Adjectives

To understand how the students described the objects, adjectives were extracted from the responses using Part-of-Speech (POS) tagging. Part-of-Speech (POS) tagging is the process of labelling each word in the sequence with corresponding part of speech (Martin & Jurafsky, 2000). For example, a tagged sentence is as below, where "AT" = Article, "VBD" = Verb in past tense, "IN" = Preposition, "NN" = Noun, "NNS" = Noun, plural.

The_AT representative_NN put_VBD chairs_NNS on_IN the_AT table_NN.

Results showed that each topic had different used adjectives, and most of them were relevant to the respective topic (Table 5.3).

Table 5.3
Top 10 adjectives of responses in Diffusion, Heat transfer, and Microfluidics

Diffusion		Heat transfer		Microfluidics	
Words	Occurrence	Words	Occurrence	Words	Occurrence
diffuse	237	caloric	491	diffuse	245
random	189	hot	385	random	89
right	155	latent	141	due	48
same	142	thermal	133	small	47
other	140	cold	92	enough	44
net	121	internal	84	turbulent	43
molecular	100	cool	82	laminar	38
high	97	solid	59	possible	38
blue	91	kinetic	59	other	35
left	89	different	41	same	35

5.1.3 Verbs

A verb represents an action or a state of an object. To understand how students used verbs in their description of processes, all verbs were extracted from the responses after POS tagging. Due to the dynamic nature of verb morphology, the stemming technique was used to reduce the words to their base form. Table 5.4 presents the occurrences of top 10 verbs used in each topic. Results showed that certain verbs, such as "move", "contain", and "flow", appeared in the top 10 for more than 1 topic. The use of these words might suggest that part of the participants were considering the phenomenon of diffusion, as substance instead of process (Slotta et al., 1995). On the contrary, "transfer" and "diffuse" were also found as the most frequently used verbs for 2 topics or more, which suggests that many participants were describing these phenomena as processes.

Table 5.4
Top 10 verbs of responses in Diffusion, Heat transfer, and Microfluidics

Diffusion		Heat transfer		Microfluidics	
Words	Occurrence	Words	Occurrence	Words	Occurrence
increase	413	transfer	261	explain	92
move	365	release	117	move	69
flow	241	flow	113	occur	67
happen	211	cool	86	include	60
contain	190	heat	82	cause	43
left	146	melt	75	diffuse	38
reach	116	change	44	continue	37
cause	111	contain	44	begin	33
occur	108	call	43	remain	31
diffuse	82	answer	41	stay	31

5.1.4 Comparison of the use of words between the control and experimental groups

To determine if there was a difference between students in control group and ontological schema trained students in their use of words, comparisons were made for each topic. Results are presented in Table 5.5, Table 5.6, and Table 5.7 for Diffusion, Heat transfer, and Microfluidics, respectively. Words in bold represent the ones that were out of the respective top 10 word list in each topic.

Under the topic "Diffusion", 9 out of 10 words were the same between control and experimental groups. The only exception was that the word "air" in control group was replaced by "random" in experimental group; both words did not appear in the top-10 word list when combining both groups. Similarly, 9 words out of the top 10 were identical between control and experimental groups in Heat Diffusion and Microfluidics. Students from the experimental group used the word "random" more frequently than the students from control group both in Diffusion and in Microfluidics.

5.2 Effectiveness of computer-based approach for identifying conceptual change

Text classification approaches were applied to identify student conceptual change based on their responses. In addition, the performances of three analytic approaches were compared, including query-based, Naïve Bayes classifier, and SVM classifier.

The training set for the classification task included 670 answers in Diffusion from session 3 of the experiment. All answers were evaluated by three different educators using the coding schema shown in Table 5.8.

Considering the limited size of training set, only two types of processes were considered, which were sequential process and emergent process. Three classification scenarios, including 2-way, 3-way, and 4-way classification, were introduced based on the two types of processes. In 2-way classification, the system was required to classify responses as either "emergent (E)" or "not emergent". In 3-way

Table 5.5
Comparison of top 10 words between control group and experimental group under the topic "Diffusion"

Control		Experimental	
Words	Occurrence	Words	Occurrence
Words	Occurrence	Words	Occurrence
water	605	molecule	735
concentrate	580	water	627
molecule	531	concentrate	487
diffuse	365	diffuse	338
increase	319	increase	327
dye	250	move	322
beaker	229	**random**	**295**
move	225	beaker	277
salt	223	dye	272
air	202	salt	213

classification, the system classifies responses as "E", "sequential (S)", or "not applicable (N/A)". In 4-way classification, the category "sequential and emergent (SE)" was added in addition to "E", "S", and "N/A"; responses under this category were those describing the complex phenomenon as both sequential and emergent processes. As shown in Table 5.9, the distribution of the labels in training set was 59 for "E", 130 for "SE", 371 for "S", and 110 for "N/A".

5.2.1 Query-based method

Keywords were used in the coding schema, so that if a response includes certain keywords, it was attributed to the respective classification category. For example, if keywords such as "random", "indirect", "continuous", "independent",

Table 5.6
Comparison of top 10 words between control group and experimental group under the topic "Heat transfer"

Control		Experimental	
Words	Occurrence	Words	Occurrence
heat	527	heat	461
energy	496	energy	419
transfer	364	transfer	289
caloric	283	caloric	253
water	211	water	211
hot	201	hot	197
ice	189	ice	196
drink	186	temperature	163
temperature	181	drink	155
flow	172	cool	150

"simultaneous", "equilibrium" appeared in a response, this response would be recognized as a description under the emergent process (E) category.

Two separate tasks using query-based approach were conducted. The first task was to retrieve responses only in "E" category; while the second task was to retrieve responses either in "E" or "SE" category (i.e., any responses related to "E"). The confusion matrices for the two tasks are shown in Table 5.10 and Table 5.11, respectively. Results showed that for the first task, 152 responses were retrieved, of which 39 and 113 responses were relevant and irrelevant to "E" category, respectively. A total of 59 responses were found under "E" category, of which 39 responses were retrieved. For the second task, 113 retrieved responses were relevant, while 39 were actually irrelevant to "E" and "SE". Fifty-six relevant responses were unable to be retrieved from the corpus.

Table 5.7
Comparison of top 10 words between control group and experimental group under the topic "Microfluidics"

Control		Experimental	
Words	Occurrence	Words	Occurrence
dye	251	diffuse	257
water	243	dye	219
diffuse	239	water	210
mix	158	flow	125
flow	134	mix	114
stream	92	stream	93
virus	84	molecule	93
molecule	81	**random**	**84**
particle	77	virus	79
bacteria	70	bacteria	75

Table 5.8
Coding schema (Miller et al., 2011)

Sequential process		Emergent process	
Direct	S-Dir	Contributes Equally	E-CE
Distinguishable	S-Dis	Moves Randomly	E-MR
Restricted	S-Res	Unintentional	E-U
Sequence	S-Seq	Indirect	E-I
Dependent	S-Dep	Equal roles	E-ER
Terminate	S-Ter	Acts simultaneously	E-AS
		Is Continuous	E-C
		Is Independent	E-Ind

Table 5.9
Distribution of the labelled responses in Diffusion for session 3

Group	Label			
	E	SE	S	N/A
Control	22	73	208	37
Experimental	37	57	163	73
Total	59	130	371	110

Table 5.10
Confusion matrix for the first task (E only) using query-based approach

		Predicted	
		Positive	Negative
Actual	Positive	39	20
	Negative	113	-

Table 5.11
Confusion matrix for the first task (E + SE) using query-based approach

		Predicted	
		Positive	Negative
Actual	Positive	113	56
	Negative	39	-

Precision, recall, and F-measure of the query-based approach were calculated from the confusion matrices (Table 5.10 and Table 5.11). Results presented in Table 5.12 showed a low precision and a fair recall when only "E" category was taken into consideration. The result is reasonable since not only responses under "E" category but also the responses under the "SE" category described the complex phenomenon as emergent processes.

For the second task, responses were retrieved in "E" or in "SE" category, and the precision reached 74% with 67% recall. This higher precision suggests that query-based approach is a simple and efficient way to identify student conceptual change using several keywords, especially when the educators rely largely on lexical features of answers when evaluating students' responses. However, this approach relies on a strict assumption that the keywords given by the educators are accurate. In addition, as the features of "sequential process" were not given, the query-based approach cannot predict whether a response belongs to the "S" category.

Table 5.12
Performance of query-based method on test set

	Precision	Recall	F_1
E only	0.26	0.66	0.37
E + SE	0.74	0.67	0.70

5.2.2 Naïve Bayes classifier

Different from the query-based approach, Naïve Bayes classifier learns features from the training set. Table 5.13 presents the results of classification performance including precision, recall, F-score and Cohen's kappa based on 10-fold cross validation. The 2-way classification was similar as the second task in query-based approach, which only considered whether the response described the phenomenon as "emergent processes". Comparing the 2-way classification of Naïve Bayes method and the second task in query-based method, the former had slightly lower precision (0.725 vs. 0.74) and recall (0.57 vs. 0.66) than the latter. In addition, the performance of Naïve Bayes method decreased as the number of categories increased, with precision declined to 0.69 and 0.61 for 3-way and 4-way classification, respectively.The Cohens kappa, which is a statistical measure for inter-rater agreement, was used to evaluate the degree of agreement between human

and machine judgements. The results showed that 2-way and 3-way classification had moderate kappa values, while 4-way classification had a poor kappa value.

Table 5.13
Performance of Naïve Bayes classifier on test set

	Precision	Recall	F_1	Kappa
2-way	0.725	0.5713	0.6331	0.5346
3-way	0.6924	0.6904	0.6804	0.4545
4-way	0.6088	0.6145	0.6029	0.3523

5.2.3 Support Vector Machine classifier

The performance of classification based on 10-fold cross validation using SVM classifier is shown in Table 5.14. It is noticeable that by using this method, the precision of 2-way classification achieved 90.95% with 64.52% recall. The SVM classifier outperformed Naïve Bayes in 2-way, 3-way and 4-way classifications, correspondingly. However, the performance of SVM classifier still decreased as the number of categories increased. The kappa value for 2-way classification using SVM method was considered as fairly good, while the 3-way and 4-way classification had moderate kappa values.

Table 5.14
Performance of SVM classifier on test set

	Precision	Recall	F_1	Kappa
2-way	0.9095	0.6452	0.7496	0.6865
3-way	0.7408	0.7298	0.7199	0.5131
4-way	0.6414	0.6494	0.634	0.4103

5.3 Evaluation of conceptual change based on computer-based approach

Based on the comparison of results of the effectiveness of different classification approaches, SVM has demonstrated better performance in the tasks. Therefore, a trained SVM classifier was used to identify whether participants demonstrated conceptual change after being exposed to the ontological schema training. The training set for this classifier contained 670 evaluated responses by human experts in topic "Diffusion" from session 3 of the experiment. The classification was then performed for answers in "Diffusion" topic from the remaining sessions 1 and 2, separately. Each classification included 2-way, 3-way, and 4-way tasks.

5.3.1 Assessment on Diffusion for Session 1

There were a total of 641 responses in Diffusion session 1. The results from 2-way classification showed that 32 out of 320 responses in control group and 97 out of 320 responses in experimental group were identified as "E", and the rest of the responses were left as "N/A".

Table 5.15 shows the results of the 3-way classification. It is noticeable that the experimental group had 96 responses as" SE+E" compared to 43 in the control group. In addition, a large portion of responses were recognized as "S", while a few were classified as "N/A", both in control and experimental groups.

The results from 4-way classification are shown in Table 5.16. Similar to the 3-way classification results, a large size of responses was classified as "S" regardless for treatments. Responses identified as "SE" in experimental group were higher compared to that in control group (41 vs. 19, respectively). Similarly, the experimental group had more responses identified as "E" compared to the control group (48 vs. 19, respectively). These results were consistent with the 2-way and 3-way classification results, indicating that there were more "E" responses in the experimental group.

Table 5.15
Results of 3-way prediction on the responses in Diffusion from session 1 using SVM classifier trained by all the evaluated responses in Diffusion from session 3

	Control	Experimental
N/A	72	36
S	205	188
SE + E	43	96
Total	320	320

Table 5.16
Results of 4-way prediction on the responses in Diffusion from session 1 using SVM classifier trained by all the evaluated responses in Diffusion from session 3

	Control	Experimental
N/A	76	41
S	206	190
SE	19	41
E	19	48
Total	320	320

Moreover, the number of responses using emergent language of each student was calculated. For 2-way and 3-way classifications, since "E" and "SE" were combined as one category, any response that either belonged to "E" or "SE" was considered as using emergent language. For 4-way classification, only responses under "E" category were considered as using emergent language. The number of questions in the test was 10, which means each student had 10 responses in total. Table 5.17 presents the descriptive statistics for number of responses using emergent language of each student.

For 2-way classification, 27 out of 32 students in the experimental group were found to use emergent language to describe the phenomena in at least one response, while the control group had 26 out of 32 students. The results showed that students in the experimental group had significant more responses using emergent language in average ($t = 5.2112$, $p < .0001$) than those in the control group. Therefore, the null hypothesis was rejected

Similarly, 28 out of 32 students in the experimental group and 27 out of 32 students in the control group were found using emergent language in at least one response in 3-way classification. Significant difference ($t = 5.7469$, $p < .0001$) was found for the average number of responses using emergent language between experimental and control groups.

In 4-way classification, "E" and "SE" were separate categories. Therefore, responses only in "E" category were considered as the case of using emergent language. 18 student in the experimental group and 23 students in the control group used emergent language in at least one response. The result was statistically significant ($t = 3.9024$, $p = .0004$), indicating that students in the experimental group responded significantly more times using emergent language in average than those in control group. A summary of the results is shown in Table 5.18.

5.3.2 Assessment on Diffusion for Session 2

The 2-way, 3-way and 4-way classifications were performed on the 610 responses from Diffusion session 2. The results of the 2-way classification showed that 50 out of 310 and 88 out of 300 responses were identified as "E" for the control and experimental groups, respectively. This agrees with the session 1 results and suggests a conceptual change of the students in experimental group.

Based on the 3-way classification, a large portion of responses were recognized as "S" in both groups (Table 5.19). Approximately 20% and 14% of the responses were classified as "N/A" in control and experimental groups, respectively.

Table 5.17
Descriptive statistics for student performance on Diffusion for session 1

Task	Group	N	Mean	Std.	Effect size
2-way					
	Experimental	27	3.5556	1.8725	1.4420
	Control	26	1.5000	0.7468	
3-way					
	Experimental	28	3.5926	1.7693	1.5609
	Control	27	1.5000	0.6814	
4-way					
	Experimental	18	2.0870	1.0999	1.2983
	Control	23	1.0556	0.2291	

Table 5.18
Results of t-test for student performance between control and experimental groups on Diffusion for session 1

Task	t	df	P-value
2-way	5.2112	51	**<.0001**
3-way	5.7469	53	**<.0001**
4-way	3.9024	39	**.0004**

Reiteratively, the experimental group had more responses answered as "E" compared to the control group.

The results from 4-way classification are shown in Table 5.20. Consistent with previous results, responses categorized as "S" account as the majority for both groups. There were 26 and 43 responses identified as "SE", and 31 and 40 were recognized as "E" in control and experimental group, respectively. Combining the "SE" and "E" responses, the results suggested that more students in the

experimental group described diffusion in an "emergent process" way compared to students in the control group.

Table 5.19
Results of 3-way prediction on the responses in Diffusion from session 2 using SVM classifier trained by all the evaluated responses in Diffusion from session 3

	Control	Experimental
N/A	63	42
S	193	168
SE + E	54	90
Total	310	300

Table 5.20
Results of 4-way prediction on the responses in Diffusion from session 2 using SVM classifier trained by all the evaluated responses in Diffusion from session 3

	Control	Experimental
N/A	63	44
S	190	173
SE	26	43
E	31	40
Total	310	300

The descriptive statistics for number of responses using emergent language in Diffusion for session 2 is shown in Table 5.21. For 2-way classification, 27 out of 31 students in the experimental group were found to use emergent language to describe the phenomena in at least one response, while the control group had 22 out of 30 students. The results showed that students in experimental group had more

responses using emergent language than those in control group, but not significant (t = 1.8474, p = .0710).

Additionally, 26 students in experimental group and 22 students in control group were found using emergent language in at least one response in 3-way classification. A significant difference (t = 2.0618, p = .0450) was found for the number of responses using emergent language of each student between experimental and control groups.

In the 4-way classification, 22 students in the experimental group and 18 students in the control group used emergent language in at least one response. The result was not statistically significant (t = 0.5086, p = .6140), indicating that students in the control and experimental groups responded almost the same times using emergent language. A summary of the results is shown in Table 5.22.

Table 5.21
Descriptive statistics for student performance on Diffusion for session 2

Task	Group	N	Mean	Std.	Effect size
2-way					
	Experimental	27	3.3704	1.9842	0.5411
	Control	22	2.4545	1.3392	
3-way					
	Experimental	26	3.4615	1.6694	0.6161
	Control	21	2.5714	1.1780	
4-way					
	Experimental	22	2.0000	1.0871	0.1627
	Control	18	1.8333	0.9574	

Table 5.22
Results of t-test for student performance between control and experimental groups on Diffusion for session 2

Task	t	df	P-value
2-way	1.8474	47	.0710
3-way	2.0618	45	**.0450**
4-way	0.5086	38	.6140

CHAPTER 6. DISCUSSION

6.1 Discussion of results

In this chapter, we will discuss the results regarding the three research goals including: 1) defining how do students describe complex phenomena with or without prior exposure to ontological schema training; 2) identifying the effectiveness of computer-based approaches for assessing student conceptual change; 3) verifying if ontological schema training can lead to student conceptual change based on computer-based assessment.

6.1.1 Use of language with and without ontological schema training

Results from the current study showed that the corpus followed the Zipf's law (Figure 5.1), indicating that the most frequently used words could represent the tendency of the words. Identification of the top 10 words used in each topic (Table 5.2) showed that different topics resulted in distinct lists of most frequently used words. This is reasonable because the given tests were highly correlated to the respective topics, of which the distribution of words were intrinsically different. Slotta et al. (1995) reported that the use of verb predicates could be used to assess conceptual change. Therefore, we further investigated the use of verbs and adjectives for all responses. Results showed that students tended to use different adjectives under different test topics, while similar verbs were used regardless of the topics (Table 5.3 and Table 5.4). One possible explanation is that the three topics tested (diffusion, heat transfer, and microfluidics) are essentially subcategories of "process", thus they share same underlying ontological attributes (Miller et al.,

2011). However, certain top 10 verbs, such as "move", "flow", "left", "contain", "include", are substance predicates instead of process predicates; this suggests that misconceptions of these topics existed among the participating students.

To evaluate how the ontological schema altered student understanding in these topics, comparison of the top 10 words for each topic were made between the control and experimental groups. Results revealed that under the topic of "diffusion" and "microfluidics", both groups used the same top 10 words with the exception of the word "random", which was only shown in the list of the experimental group (Table 5.5 and Table 5.7). The word "random" is one of the keywords that is used to identify emergent process descriptions during traditional manual assessment by educators (Miller et al., 2011). The high occurrence of "random" in both groups may indicate that students accurately described these two phenomena as emergent processes. Therefore, the ontological schema training may have helped students in the experimental group to better understand and interpret the phenomena being tested. This is consistent with Yang et al. (2012), who reported that the integration of ontological schema training into online learning modules could help students understand complex phenomena better. Similarly, Streveler et al. (2013) also showed improved understandings among students who were exposed to schema training that used more emergent language than those without the training.

6.1.2 Effectiveness of computer-based approach for identifying conceptual change

The performance of three different text classification methods were compared on classifying the evaluated responses from session 3 under the topic of "diffusion". Although the query-based approach showed good performance ($F_1 = 0.70$, Table 5.12), the main drawback of this method is its dependence on accurate keywords, which are provided by educators to identify the correct answers. This dependency may result in undervaluation of students responses if other morphology of the

keywords should be used. To solve this problem, a supervised learning method can be applied, which can allow the system to learn features from the training dataset. In the current study, two supervised learning methods, including the Naïve Bayes and SVM methods, were introduced. Naïve Bayes classifier has been proved to be an efficient method in classification tasks, therefore it was chosen to be the baseline. However, Naïve Bayes method holds a strong assumption that the features are conditionally independent, indicating that information about relationship between features is discarded. The results showed that both Naïve Bayes and SVM classifiers had higher accuracy in identifying student conceptual change from the responses as compared to the query-based approach. In addition, the SVM classifier outperformed Naïve Bayes classifier in 2-way, 3-way, and 4-way classifications (Table 5.13 and Table 5.14).

Previous research suggests that assessment of student responses needs to exploit deeper semantic features more than shallow features (Dzikovska et al., 2013). Therefore, semantic features were applied with different supervised learning methods to assess student responses (Aldabe, Maritxalar, & de Lacalle, 2013; Okoye, Bethard, & Sumner, 2013; Zesch, Levy, Gurevych, & Dagan, 2013). The results are consistent with previous research though only lexical features were used. One possible explanation is that there was a clear difference of the use of words between sequential and emergent language making it classifiable.

6.1.3 Evaluation of student conceptual change based on computer-based approach

Based on the trained SVM classifier, three different classification tasks (i.e., 2-way, 3-way, and 4-way classification) were performed on responses from session 1 and 2 under the topic of "Diffusion".

The 2-way classification results showed that students in the experimental group used emergent language more often than those in the control group. The results of statistical analysis (Table 5.18) suggested that students in the

experimental group had significant more responses using emergent language ($t = 5.2112$, $p < .0001$) than those in the control group. This finding is consistent with Streveler et al. (2013), who reported that ontological schema training led to significantly higher usage of emergent language. However, the results from 3-way classification showed that a large proportion of responses (Control 64.06%, Experimental 58.75%) mistakenly described the phenomena (diffusion) as sequential processes (Table 5.15). Similarly, Streveler et al. (2013) found that the use of emergent language was not predominant regardless of ontological schema training. The authors suggested that because the participants had finished at least one course in the tested topic (heat transfer) prior to the experiment, thus conceptual change might be difficult to occur in such a short time, if they had already formed misconceptions from earlier courses (Yang et al., 2012). Moreover, results from 4-way classification (Table 5.16) suggested that half of the responses recognized as "E" also used sequential language ("SE" responses = 19, "SE+E" responses = 43). This again suggests the difficulty of repairing misconceptions through ontological schema training upon a short period of time; although ontological schema training has previously shown to be effective (Streveler et al., 2013).

6.2 Implications for Educational Research

Conceptual change is an important approach to identify and repair misconceptions in science and engineering. One way of measuring conceptual change is to assess the use of verb predicates in student responses (Slotta et al., 1995), which can be done by identifying how students describe concepts in scientifically correct language (Streveler et al., 2013). Instead of manually analyzing student responses, computer-based approaches provide an automatic and more effective way to assess student conceptual change.

The results of this study suggest that conceptual change identification can be integrated with online learning modules; this integration makes it possible to

provide students immediate feedback once a misconception has been identified. Feedback can greatly influence learning and teaching. From the perspective of learning, immediate feedback assessment technique (IFAT) has shown to be an effective method that engages learners in the discovery processes, during which it promotes retention, and helps students to correct initial inaccurate responses (Epstein et al., 2010). From the perspective of teaching, Hattie and Timperley (2007) pointed out that "teachers need to seek and learn from student responses to test as much as do students" (p. 104). Leveraging on computer-based assessment approaches, teachers can immediately know how the students learn.

CHAPTER 7. CONCLUSION AND FUTURE WORK

Current study provides future opportunities to investigate student conceptual change in other learning environments, such as distance learning and massive open online courses (MOOC). With rapid growth of student enrolment in MOOCs, it is impossible to use traditional manual assessment approaches. Current assessments in MOOCs include peer grading, automatic grading for essays (Balfour, 2013), and programming assignments (Pieterse, 2013). The development of automatic methods for assessing student conceptual change can be a promising direction for fast and effective assessment in MOOCs and distance learning.

Due to the fact that only responses in topic "Diffusion" were evaluated manually, we were unable to evaluate the computer-based approaches across the three topics. However, the first finding, which showed there was a difference of use of language between control and experimental groups, despite of the topic, suggests that the results may also apply for other topics. A limitation is that the classifications were performed in a document level instead of a phrase level. In addition, the length of responses and unbalanced dataset may also be limitations for the classification performance in this study. Future work include (a) the generalization of computer-based assessments mechanisms for conceptual change in different domains (b) improvement of the performance of conceptual change identification, and (c) integration of conceptual change assessment into online learning modules to provide feedback for students.

LIST OF REFERENCES

LIST OF REFERENCES

Aldabe, I., Maritxalar, M., & de Lacalle, O. L. (2013). Ehu-alm: Similarity-feature based approach for student response analysis. In *Second joint conference on lexical and computational semantics (* sem).*

Balfour, S. P. (2013). Assessing writing in moocs: automated essay scoring and calibrated peer review. *Research & Practice in Assessment*, *8*(1), 40–48.

Bienkowski, M., Feng, M., & Means, B. (2012, Oct.). Enhancing teaching and learning through educational data mining and learning analytics: An issue brief.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT press.

Chi, M. T. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *The Journal of the Learning Sciences*, *14*(2), 161–199.

Chi, M. T., & Roscoe, R. D. (2002). The processes and challenges of conceptual change. In *Reconsidering conceptual change: Issues in theory and practice* (pp. 3–27). Springer.

Chi, M. T., Roscoe, R. D., Slotta, J. D., Roy, M., & Chase, C. C. (2012). Misconceived causal explanations for emergent processes. *Cognitive science*, *36*(1), 1–61.

Chi, M. T., Slotta, J. D., & De Leeuw, N. (1994). From things to processes: a theory of conceptual change for learning science concepts. *Learning and Instruction*, *4*(1), 27–43.

Deerwester, S. C., Dumais, S. T., Landauer, T., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on human language technology research* (pp. 138–145).

Dzikovska, M. O., Nielsen, R. D., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., ... Dang, H. T. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second joint conference on lexical and computational semantics (* sem)* (Vol. 2, pp. 263–274).

Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., & Brosvic, G. M. (2010). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *The Psychological Record*, *52*(2), 5.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81–112.

He, Y., Hui, S. C., & Quan, T. T. (2009). Automatic summary assessment for intelligent tutoring systems. *Computers & Education*, *53*(3), 890–899.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*, 141.

Landauer, T., & Dutnais, S. T. (1997). A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 211–240.

Landauer, T., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2-3), 259–284.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. (1997). How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. In *Proceedings of the 19th annual meeting of the cognitive science society* (pp. 412–417).

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 conference of the north american chapter of the association for computational linguistics on human language technology-volume 1* (pp. 71–78).

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1). Cambridge University Press Cambridge.

Martin, J. H., & Jurafsky, D. (2000). *Speech and language processing.* Prentice hall.

Miller, R. L., Streveler, R. A., Nelson, M. A., Geist, M. R., & Olds, B. M. (2005). Concept inventories meet cognitive psychology: Using beta testing as a mechanism for identifying engineering student misconceptions. In *Proceedings of the american society for engineering education annual conference* (pp. 12–15).

Miller, R. L., Streveler, R. A., Yang, D., & Santiago Román, A. I. (2011). Identifying and repairing student misconceptions in thermal and transport science: Concept inventories and schema training studies. *Chemical Engineering Education*, *45*(3), 203–210.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 752–762).

Nielsen, R. D., Ward, W., & Martin, J. H. (2008). Learning to assess low-level conceptual understanding. In *Flairs conference* (pp. 427–432).

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified bleu algorithm. *Computers & Education*, *56*(2), 337–345.

Okoye, I., Bethard, S., & Sumner, T. (2013). Cu: Computational assessment of short free text answers-a tool for evaluating students understanding. In *Second joint conference on lexical and computational semantics (* sem).*

Olde, B. A., & Franceschetti, D. R. (2002). The right stuff: do you need to sanitize your corpus when using latent semantic analysis? In *Proceedings of the 24th annual meeting of the cogonitive science society* (pp. 708–713).

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311–318).

Pérez, D., Alfonseca, E., & Rodríguez, P. (2004). Application of the bleu method for evaluating free-text answers in an e-learning environment. In (pp. 26–28).

Pérez, D., Gliozzo, A., Strapparava, C., Alfonseca, E., Rodríguez, P., & Magnini, B. (2005). Automatic assessment of students' free-text answers underpinned by the combination of a bleu-inspired algorithm and latent semantic analysis. In *Proceedings of the eighteenth international florida artificial intelligence research society conference* (pp. 358–362).

Petersen, S. E., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech & Language*, *23*(1), 89–106.

Pieterse, V. (2013). Automated assessment of programming assignments. In *Proceedings of the 3rd computer science education research conference on computer science education research* (pp. 45–56).

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: the role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, *63*(2), 167–199.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Proceedings of the second workshop on building educational applications using nlp* (pp. 9–16).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Siemens, G., & Baker, R. S. d. (2012). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252–254).

Slotta, J. D., & Chi, M. T. (2006). Helping students understand challenging topics in science through ontology training. *Cognition and Instruction*, *24*(2), 261–289.

Slotta, J. D., Chi, M. T., & Joram, E. (1995). Assessing students' misclassifications of physics concepts: An ontological basis for conceptual change. *Cognition and Instruction*, *13*(3), 373–400.

Society for learning analytics research. (2011, Oct.). Retrieved from http://www.solaresearch.org/

Streveler, R. A., Litzinger, T. A., Miller, R. L., & Steif, P. S. (2008). Learning conceptual knowledge in the engineering sciences: overview and future research directions. *Journal of Engineering Education*, *97*(3), 279–294.

Streveler, R. A., Miller, R. L., Perova-Mello, N., Pitterson, N., Denick, D., Magana, A. J., ... Fayyaz, F. (2013). Can "emergent" language serve as an indicator of conceptual change? In *Proceedings of the 15th biennial earli conference for research on learning and instruction.*

Strike, K. A., & Posner, G. J. (1992). A revisionist theory of conceptual change. *Philosophy of Science, Cognitive Psychology, and Educational Theory and Practice*, 147–176.

Thelwall, M. (2000). Computer-based assessment: a versatile educational tool. *Computers & Education*, *34*(1), 37–49.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: a study of conceptual change in childhood. *Cognitive Psychology*, *24*(4), 535–585.

Yang, D., Streveler, R. A., Miller, R. L., Slotta, J. D., Matusovich, H. M., & Magana, A. J. (2012). Using computer-based online learning modules to promote conceptual change: helping students understand difficult concepts in thermal and transport science. *International Journal of Engineering Education*, *3*, 686-700.

Zesch, T., Levy, O., Gurevych, I., & Dagan, I. (2013). Ukp-biu: Similarity and entailment metrics for student response analysis. In *Second joint conference on lexical and computational semantics (* sem).*

APPENDIX

APPENDIX. APPROVAL FROM INSTITUTIONAL REVIEW BOARD

Revised 10/10 Ref. #

APPLICATION TO USE HUMAN RESEARCH SUBJECTS

Purdue University
Institutional Review Board

1. Project Title: Automatic detection of students' misconceptions in thermal and transport science

2. Full Review ☐ Expedited Review ☒

3. Anticipated Funding Source: N/A

4. Principal Investigator *[* See Policy on Eligibility to serve as a Principal Investigator for Research Involving Human Subjects*]*:

Name and Title	Department, Building, Phone, FAX, E-mail address
Alejandra J. Magana, Assistant Professor	CIT, KNOY 231, 494-3994, , admagana@purdue.edu

5. Co-investigators and key personnel *[See Education Policy for Conducting Human Subjects Research]*:

Name and Title	Department, Building, Phone, FAX, E-mail address
Junchao Yan, Graduate Student	CIT, ,765-337-8867, ,yan114@purdue.edu

6. Consultants *[See Education Policy for Conducting Human Subjects Research]*:

Name and Title	Department, Building, Phone, FAX, E-mail address

7. The principal investigator agrees to carry out the proposed project as stated in the application and to promptly report to the Institutional Review Board any proposed changes and/or unanticipated problems involving risks to subjects or others participating in the approved project in accordance with the HRPP Guideline 207 Researcher Responsibilities, Purdue Research Foundation-Purdue University Statement of Principles and the Confidentiality Statement. The principal investigator has received a copy of the Federal-Wide Assurance (FWA) and has access to copies of 45 CFR 46 and the Belmont Report. The principal investigator agrees to inform the Institutional Review Board and complete all necessary reports should the principal investigator terminate University association.

Principal Investigator Signature _____ Date _____

8. The Department Head (or authorized agent) has read and approved the application. S/he affirms that the use of human subjects in this project is relevant to answer the research question being asked and has scientific or scholarly merit. Additionally s/he agrees to maintain research records in accordance with the IRB's research records retention requirement should the principal investigator terminate association with the University.

Department Head *(printed)* _____ Department Name _____

Department Head Signature _____ Date _____

APPLICATION TO USE HUMAN RESEARCH SUBJECTS

9. This project will be conducted at the following location(s): (please indicate city & state)
   - ☒ Purdue West Lafayette Campus
   - ☐ Purdue Regional Campus (Specify): _____
   - ☐ Other (Specify): _____

10. If this project will involve potentially vulnerable subject populations, please check all that apply.
   - ☐ Minors under age 18
   - ☐ Pregnant Women
   - ☐ Fetus/fetal tissue
   - ☐ Prisoners Or Incarcerated Individuals
   - ☐ University Students (PSYC Dept. subject pool ___)
   - ☐ Elderly Persons
   - ☐ Economically/Educationally Disadvantaged Persons
   - ☐ Mentally/Emotionally/Developmentally Disabled Persons
   - ☐ Minority Groups and/or Non-English Speakers
   - ☐ Intervention(s) that include medical or psychological treatment

11. Indicate the anticipated maximum number of subjects to be enrolled in this protocol as justified by the hypothesis and study procedures: 100

12. This project involves the use of an **Investigational New Drug** (IND) or an **Approved Drug For An Unapproved Use**.
   ☐ YES ☒ NO
   Drug name, IND number and company: _____

13. This project involves the use of an **Investigational Medical Device** or an **Approved Medical Device For An Unapproved Use**.
   ☐ YES ☒ NO
   Device name, IDE number and company: _____

14. The project involves the use of Radiation or Radioisotopes:
   ☐ YES ☒ NO

15. Does this project call for: (check-mark all that apply to this study)
   - ☐ Use of Voice, Video, Digital, or Image Recordings?
   - ☐ Subject Compensation? Please indicate the maximum payment amount to subjects. $_____
     Purdue's Human Subjects Payment Policy Participant Payment Disclosure Form
   - ☐ VO2 Max Exercise?
   - ☐ More Than Minimal Risk?
   - ☒ Waiver of Informed Consent?
   - ☐ Extra Costs To Subjects?
   - ☐ The Use of Blood? Total Amount of Blood _____ Over Time Period (days) _____
   - ☐ The Use of rDNA or Biohazardous materials?
   - ☐ The Use of Human Tissue or Cell Lines?
   - ☐ The Use of Other Fluids that Could Mask the Presence of Blood (Including Urine and Feces)?
   - ☐ The Use of Protected Health Information (Obtained from Healthcare Practitioners or Institutions)?
   - ☐ The Use of academic records?

16. Does investigator or key personnel have a potential financial or other conflict of interest in this study?
   ☐ YES ☒ NO

APPLICATION NARRATIVE

A. PROPOSED RESEARCH RATIONALE

- Describe why you are conducting the study. Identify the research question being asked.

The PI of this study is requesting a waiver for informed consent to use the data from the previous project supported by National Science Foundation through grant EEC-0550169, 'Developing Ontological Schema Training Methods to Help Students Develop Scientifically Accurate Mental Models of Engineering Concepts'. This project applies Natural Language Processing and Machine Learning techniques to automatically detect misconceptions in participants' responses (data from the previous project) in thermal and transport science. This technique will free people from the need of manually organizing large sets of documents and make the procedure for identifying misconceptions more effectively.

Research questions:
1. Can text classification (TC) technique be used for categorization of students' misconceptions in thermal and transport science?
2. What is the effectiveness of using TC for identifying substance and emergent students written explanations of thermal and transport science?

B. SPECIFIC PROCEDURES TO BE FOLLOWED

- Describe in a step-by-step manner what you will require subjects to do in this study.

Subjects will not have to do anything for this study. The data has already been collected.

- Identify all data you will collect.

The data includes users' login ID and open-ended responses to the questions probing their explanations of their understanding of thermal and transport science phenomena

C. SUBJECTS TO BE INCLUDED

Describe:

- The inclusion criteria for the subject populations including gender, age ranges, ethnic background, health status and any other applicable information. Provide a rationale for targeting those populations.
- The exclusion criteria for subjects.
- Explain the rationale for the involvement of any special populations including prisoners
- Provide the maximum number of subjects you seek approval to enroll from all of the subject populations you intend to use and justify the sample size. You will not be approved to enroll a number greater than this. If at a later time it becomes apparent you need to increase your sample size, you will need to submit a Revision Request.
- **For NIH funded protocols**: If you do not include women, minorities and children in your subject pool, you must include a justification for their exclusion. The justification must meet the exclusionary criteria established by the NIH.

This study does not require any active recruitment of subjects or any direct contact with participants. We are requesting permission to use the data from previous project and apply automatic (computer based) classification methods to identify participants' misconceptions in thermal and transport science. The population of users is unknown.

D. RECRUITMENT OF SUBJECTS AND OBTAINING INFORMED CONSENT

- Describe your recruitment process in a step-by-step manner. The IRB needs to know all the steps you will take to recruit subjects in order to ensure subjects are properly informed and are participating in a voluntary manner. An incomplete description will cause a delay in the approval of your protocol application.

Participants will not be recruited as part of this study. We are requesting permission to use the data from previous project and apply automatic classification methods to identify participants' misconceptions in thermal and transport science. The user login ID is connected to the response in each single observation, and no other identifiable data will be used as part of this study. Furthermore, user login ID will only be used as an internal identifier of the data. The results of this study will be reported in aggregated form where no user login ID will be connected to the data.

E. PROCEDURES FOR PAYMENT OF SUBJECTS

- Describe any compensation that subjects will receive. Please note that Purdue University Business Services policies might affect how you can compensate subjects. Please contact your department's business office to ensure your compensation procedures are allowable by these policies.

N/A

F. CONFIDENTIALITY

- Describe what steps you will take to maintain the confidentiality of subjects.
- Describe how research records, data, specimens, etc. will be stored and for how long. The IRB generally recommends locked storage, such as a cabinet, for identifiable information. Please note, consent forms signed by subjects, parents and/or legally authorized representatives ARE considered research records.
- Describe if the research records, data, specimens, etc. will be de-identified and/or destroyed at a certain time. If records, data, specimens, etc. will be de-identified, address if a code key will be maintained and when, if ever, it will be destroyed. Additionally, address if they may be used for future research purposes.

To maintain the confidentiality of the subjects, the user login ID of every user will be replaced by an internal identifier. Furthermore, findings of this study will be reported in aggregated form where no user login ID will be connected to the data.

G. POTENTIAL RISKS TO SUBJECTS

- There are always risks associated with research. If the research is minimal risk, which is no greater than every day activities, then please describe this fact.
- Describe the risks to participants and steps that will be taken to minimize those risks. Risks can be physical, psychological, economic, social, legal, etc.
- Where appropriate, describe alternative procedures or treatments that might be advantageous to the participants.
- Describe provisions for ensuring necessary medical or professional intervention in the event of adverse effects to participants or additional resources for participants.

The risk for this project is minimal and no greater than every day activities. A potential risk inherent to any technological tool that requires from the user personal data is breach of information. This risk could happen regardless of whether or not we conduct the proposed study. However, all measures will be taken to avoid breach of confidentiality of the data. For example, user login ID will be replaced with another internal identifier and findings resulted from this study will be reported in aggregated form. Risks greater than those encountered in everyday life are not anticipated.

H. BENEFITS TO BE GAINED BY THE INDIVIDUAL AND/OR SOCIETY

- Describe the possible direct benefits to the subjects. If there are no direct benefits, please state this fact.
- Describe the possible benefits to society.

There is no direct benefit to the subjects. However, the study of applying automatic technique will free people from the need of manually organizing large sets of documents and make the procedure for identifying misconceptions in science and engineering more effective. Moreover, the scientific community will have better understanding of applying Natural Language Processing and Machine Learning techniques for educational research data analysis.

I. INVESTIGATOR'S EVALUATION OF THE RISK-BENEFIT RATIO

The risk to benefit ratio for this study is extremely low. The greatest benefit of conducting this study is to provide better technological tools to researchers, educators and learners. The greatest potential risk with participating in the study would be potential breach of identity with being identified as a participant. This risk is far out-weighed by the direct and indirect benefits mentioned above.

J. WRITTEN INFORMED CONSENT FORM ***(to be attached to the Application Narrative)***

- Submit a copy of the informed consent document in the form that it will be disseminated to subjects. The approved consent form will be stamped with the IRB's approval and returned to you for use.
- If recruiting subjects who do not speak English, submit both an English version as well as a version translated into the appropriate foreign language.

N/A

K. WAIVER OF INFORMED CONSENT OR SIGNED CONSENT

If requesting either a waiver of consent or a waiver of signed consent, please address the following:

1. For a Waiver of Consent Request, address the following:
   a. Does the research pose greater than minimal risk to subjects (greater than everyday activities)?

   No, the research does not pose greater than minimal risks to subjects.

   b. Will the waiver adversely affect subjects' rights and welfare? Please justify?

   No, the waiver will not adversely affect subjects' rights.

   c. Why would the research be impracticable without the waiver?

   It is impractical and almost impossible to contact all users again to obtain consent. Also, the study will not be reliable and will not represent any finding if only part of data is used and analyzed.

   d. How will pertinent information be reported to subjects, if appropriate, at a later date?

   Information will not be directly reported to subjects. The results of this study will be published as a scientific paper.

2. For a Waiver of Signed Consent, address the following:
   a. Does the research pose greater than minimal risk to subjects (greater than everyday activities)?
   b. Does a breech of confidentiality constitute the principal risk to subjects?
   c. Would the signed consent form be the only record linking the subject and the research?
   d. Does the research include any activities that would require signed consent in a non-research context?
   e. Will you provide the subjects with a written statement about the research (an information sheet that contains all the elements of the consent form but without the signature lines)?

L. INTERNATIONAL RESEARCH

When conducting international research investigators must provide additional information to assist the IRB in making an appropriate risk/benefit analysis. Please consult the bullet points below when addressing this section of the application.

- Research projects must be approved by the local equivalent of an IRB before Purdue's IRB can grant approval to the protocol. If there is not equivalent board or group, investigators must rely on local or cultural experts or community leaders to provide approval and affirm the research procedures are appropriate for that culture. The Purdue IRB requires documentation to be submitted of this "local approval" before granting approval of the protocol. Additionally, please provide information about the IRB equivalent and provide contact information for the local entity. The body or individual providing the local approval should be identified in the application narrative as well as information as to that body's or individual's expertise.
- In the application narrative describe the experience and/or other qualifications the investigators have related to conducting the research with the local community/culture. Describe if the investigators have the knowledge or expertise of the local or state or national laws that may impact the research. The investigators must understand community/cultural attitudes to appreciate the local laws, regulations or norms to ensure the research is conducted in accordance with U.S. regulations as well as local requirements.
- For more information on specific requirements of different countries and territories, investigators can consult the Office for Human Research Protections International Compilation of Human Research Protections (http://www.hhs.gov/ohrp/international/). This is only one resource and it may not be an appropriate resource for your individual project.
- In the application narrative describe how the investigators will have culturally appropriate access to the community. If the investigators were invited into the community to conduct the research, please submit documentation of the collaboration.
- In the application narrative explain the investigators' ability to speak, read or write the language of potential participants. Describe the primary language spoken in the community. Explain provisions for culturally appropriate recruitment and consent accommodations translated materials or translators.
- Attention should be given to local customs as well as local cultural and religious norms when writing consent documents or proposing alternative consent procedures. This information should be provided in the application narrative, and as appropriate, provide justification if requesting the IRB to waive some or all requirements of written consent.
- In the application narrative describe how investigators will communicate with the IRB while you are conducting the research in the event the project requires changes or there are reportable events. Also, if the researcher is a student, describe how the student will communicate with the principal investigator during the conduct of the research and how the principal investigator will oversee the research.
- If this research is federally funded by the United States, additional documentation and inter-institutional agreements may be required. Contact the IRB Administrator for assistance.
- Submit copies of consent documents and any other materials that will be provided to subjects (e.g., study instruments, advertisements, etc.) in both English and translated to any other applicable languages.

N/A

M. SUPPORTING DOCUMENTS ***(to be attached to the Application Narrative)***

- Recruitment advertisements, flyers and letters.
- Survey instruments, questionnaires, tests, debriefing information, etc.
- If the research is a collaboration with another institution, the institution's IRB or ethical board approval for the research.
- If the research accesses the PSYC 120 Subject pool include the description to be posted on the web-based recruitment program (formerly *Experimetrix*).
- Local review approval or affirmation of appropriateness for international research.

- If the research will be conducted in schools, businesses or organizations, include a letter from an appropriate administrator or official permitting the conduct of the research.

N/A