Kim, M., Zouaq, A., & Kim, S. (2016). Automatic detection of expert models: The exploration of expert modeling methods applicable to technology-based assessment and instruction. *Computers & Education*, *101*, 55-69.

Automatic Detection of Expert Models: The Exploration of Expert Modeling Methods Applicable to Technology-Based Assessment and Instruction

Abstract

This mixed methods study explores automatic methods for expert model construction using multiple textual explanations of a problem situation. In particular, this study focuses on the key concepts of an expert model. While an expert understanding of a complex problem situation provides critical reference points for evidence-based formative assessment and feedback, the extraction of those reference points has proven challenging. Building upon semantic analysis, this study utilizes deep natural language processing techniques to facilitate the automatic extraction of key concepts from textual explanations written by experts. The study addresses the following question: (a) whether experts in a domain represent a common understanding of a problem situation through shared key concepts, (b) which metrics extract key concepts from textual data most accurately, and (c) whether automatic methods enable expert model construction from a corpus of textual explanations instead of a pre-defined, ideal explanation created using the Delphi method. The OntoCmap tool was used to extract concepts from multiple textual explanations and to generate diverse metrics assigned to each concept. The findings indicate that (a) experts have varying ways of understanding a problem situation, (b) graph-based filtering metrics (i.e., betweenness and reachability) performed better in building a set of key concepts, and (c) a single, pre-defined explanation led to a more accurate set of key concepts than a corpus of explanations from various experts.

Keywords: Expert models, natural language processing, problem solving, graph-based metrics, formative assessment

Automatic Detection of Expert Models: The Exploration of Expert Modeling Methods Applicable to Technology-Based Assessment and Instruction

1. Introduction

Expert studies have sought to explain what constitutes expertise in a particular task and how expertise develops across short-term and long-term changes (Alexander, 2004; Chi, 2006; Ericsson, Charness, Feltovich, & Hoffman, 2006; Flavell, 1992). Experts exhibit knowledge and skills in problem solving that are clearly distinguishable from novices (Glaser, Chi, & Farr, 1988; Spector & Koszalka, 2004). Expert models have been used not only to shape curriculum (Nkambou, 2010) and instructional materials (Gobet & Wood, 1999; Seel, 2003) but also to scaffold learning progressions (Feyzi-Behnagh & Azevedo, 2012; Molloy & Boud, 2014). For example, the key concepts provided by expert models are vital to evidence-based assessment, precise and targeted feedback, and learning guidance (XXX, 2015; Shute et al., 2010). Among the many aspects of an expert model, the current study focuses on expert-identified key concepts that explain a particular problem situation.

Key concepts identified by experts are often used in educational learning systems. For example, Intelligent Tutoring Systems (ITSs) such as AutoTutor use pre-defined key concepts to detect student understanding of a problem through dialogue with a system agent and to provide higher-level feedback such as prompts, hints, and examples (Lintean, Rus, & Azevedo, 2012; Rus, D'Mello, Hu, & Graesser, 2013). Suppose that students in an earth science class are asked to write responses to a system-generated question: "describe which factors have influenced the frequency of wildfires in California in 2015?" An agent could identify concepts from student responses and compare them to the concepts in an expert model. Another popular application is automated essay scoring to evaluate written work. For instance, the c-rater scoring engine developed by ETS is designed to recognize specific content (e.g., concepts) in a student response (Burstein, Tetreault, & Madnani, 2013; Gopal et al., 2010; Shermis, 2010; Shermis, Burstein, & Leacock, 2006). Automated natural language processing techniques determine whether a student response contains information that shows a topic has been learned (Shermis, 2010; Shermis, Burstein, & Leacock, 2006).

In spite of all these benefits, modeling expert knowledge is demanding, expensive, and time consuming. For instance, building a consentient expert model of a complex problem situation often involves diverse and inter-related concepts, principles, propositions, and various unknown features (Jonnassen & Philip, 1996; XXX, 2012; Spector, 2010). In order to tackle these issues, current technologies have promoted methods for automatically representing expert knowledge that potentially contains the key concepts of a problem situation (Clariana, Wolfe, & Kim, 2014; XXX, 2013).

One of the most popular methods involves natural language processing (NLP), which can be used to analyze textual explanations in order to construct knowledge models (Allen, Snow, & McNamara, 2015). For example, many ITSs use NLP tools to analyze the linguistic properties of conversations between an agent and learner, a process that results in student models (McNamara, Crossley, & Roscoe, 2013; Nye, Graesser, & Hu, 2014; Rus et al., 2013). The same techniques can be used to build expert models, against which student models are compared in order to identify knowledge gaps. Yet whether these technological methods can generate expert concepts as reliably and efficiently as human can remains debatable (Allen et al., 2015). In an attempt to respond to these issues, the current study explores automatic methods for extracting key concepts from multiple expert explanations of a complex problem situation.

2. Theoretical and Methodological Underpinnings

2.1. Expert modeling for a complex problem situation

Expert models are essential to an accurate evaluation of student input in technology-enhanced learning systems. Nevertheless, whether building an expert model is always plausible or reasonable remains an open question. In fact, one might experience difficulty in establishing reference models for a complex problem situation. Complex problems (e.g., an ecology problem in 9th-grade science class—identifying the complex nature of global warming and its connection to frequent wild fires in California) involve interrelated and multilayered concepts, factors, and unknown variables, which often render a common understanding of the problem situation impossible (Jonassen, 2014; Pretz, Naples, & Sternberg, 2003).

Despite this skepticism, we still need expert models to identify, monitor, and promote levels of expertise in problem-solving situations. Anderson (1980) defined problem solving as a goal-directed cognitive effort. In the course of problem solving, problem solvers typically conceptualize the problem space in which all of the possible conditions of a problem exist (Newell & Simon, 1972; Pretz et al., 2003). Although experts in the same discipline might exhibit a similar and clearly recognizable understanding of a problem situation, their solutions could be multiple and diverse (Spector, 2008). Therefore, the current study analyzed expert conceptualizations—problem spaces— of a problem situation and investigated how similar or dissimilar these problem spaces were. To this end, natural language (e.g., written responses) is often used to elicit expert understanding for two primary reasons: it enables descriptive verbal knowledge representations (Axelrod, 2015; Pirnay-Dummer & Ifenthaler, 2010) and saves time and effort (Brown, 1997). Cognitive and learning scientists have used verbalized explanations to model expert knowledge(Allen et al., 2015; Chi, De Leeuw, Chiu, & Lavancher, 1994; Dascalu, Trausan-Matu, Dessus, & McNamara, 2015; Johnson-Laird, 2013). Externalized articulation reflects internal mental structures. Accordingly, one can infer internal mental models by extracting meaningful concepts and their relations (i.e., propositions) from language representations (e.g., expert accounts of a problem) (R. B. Clariana & Wallace, 2009; Jonassen, Beissner, & Yacci, 1993).

2.2. Source of expertise modeling

The general approach to constructing an expert model (i.e., key concepts of a problem situation) is to draw on human expert-generated data. Lintean et al. (2012) proposed three sources used by intelligent learning systems to obtain human expert data related to a problem situation: (a) a set of key concepts related to the problem task directly defined by human experts, (b) existing text-based references (e.g., book chapters) endorsed by human experts, and (c) ideal or expected explanations of a problem situation composed by human experts.

Key concept development through direct human judgment is typically labor-intensive and time-consuming and does not easily yield a list of potential key concepts. For text-based sources, NLP and statistical algorithms are necessary to extract key concepts. The second source type, an existing text-based reference, is only available when a document happens to be exactly related to a problem task. In many educational contexts related to particular learning objectives, finding existing reference documents that fit the context of a problem situation is rare. For this reason, the current study relied on the third source type, an ideal or expected textual explanation of a given problem that is likely to reflect the thought process of an expert. We also contrasted individual expert explanations with the ideal explanation to identify any common views of a given problem.

2.3. Concept map-based approach

While scholars agree that language plays a critical role in representing expert models, there are different theoretical and methodological approaches to analyzing textual data and identifying key concepts. For example, traditional information retrieval techniques have used statistical measures of frequency of occurrence for term extraction, such as Term Frequency–Inverse Document Frequency (TF-IDF), typically using a large set of linguistic corpora (Manning, Raghavan, & Schütze, 2008; Salton & Buckley, 1988). Other approaches build on concept maps to elicit cognitive representations of relevant knowledge structures from textual responses. Through these approaches, a concept map elicited from an expert textual explanation can represent an expert knowledge structure for a particular problem task (Coronges, Stacy, & Valente, 2007).

Key concepts are assumed to emerge from a concept map in which all of the ideas in a text are interrelated in an array of concept-to-concept relations (R. B. Clariana et al., 2014; Narayanan, 2005; Novak & Cañas, 2006). Therefore, the identification of key concepts necessitates the construction of a reliable concept map as a semantic structure. Approaches to constructing concept maps can be characterized as ways of identifying concepts and determining the relations (i.e., propositions) among those concepts. For example, building upon the assumption that concepts (e.g., nouns) with meaningful relations are located near each other in a textual space, a simple approach considers that two concepts adjacent to each other are related (R. Clariana, Wallace, & Godshalk, 2009). According to another approach, a pair of concepts that are closer to each other are regarded as having stronger associations. Therefore, each pair of concept-to-concept relations is weighed with the average number of words between the two concepts, and these weighted values are used to identify key relations that feature key concepts (Pirnay-Dummer, Ifenthaler, & Spector, 2009). These two approaches are methodologically simple and technically convenient. However, these shallow NLP approaches do not attempt to understand the exact meanings of word pairs in a text that build a concept map or determine the relative significance of concepts in a concept map; therefore, they often ignore many aspects of textual argumentation. They sometimes even produce insufficient or erroneous knowledge representations and lead to improper interpretation (XXX, 2012; Shaffer et al., 2009; Xu & Krieger, 2003).

In contrast to these shallow NLP approaches, the current study uses deep NLP techniques to construct a more meaningful concept map (i.e., a network of semantic relations) from which more accurate key concepts are likely to be extracted (see Figure 1). For example, XXX (2012) proposed the semantic relation approach, which argues that richer knowledge structures can be modeled by extracting semantic relations that logically correspond with the associated elements of a sentence (Girju et al., 2009; Zouaq, Gasevic, & Hatala, 2011b). The underlying assumption of the semantic relation approach is that the logical connection (i.e., semantic relation) in a concept pair is determined by several syntactic patterns (Janssen, 2012; Montague, 1974). For example, to build an expert model, XXX (2013) identified concept pairs with semantic relations that were determined by phrase-level patterns (e.g., *the library of the school*; the logical sense is that the library belongs to the school) or sentence-level patterns (e.g., *the school has a new technology*). Indeed, a cross-validation study demonstrated that elicited knowledge structures and key concepts varied quite extensively across applied methods (XXX, 2012). In addition, XXX (2011) defined patterns through syntactic analysis (a) to identify candidate concepts (e.g., nouns, noun compounds, and nouns modified by adjectives) and (b) to extract labeled relations between candidate concepts using grammatical dependencies. These relations resulted in the creation of concept maps around candidate concepts.

— Insert Figure 1 here —

In this regard, the current study used a domain ontology tool, OntoCmap (XXX, 2014; XXX, 2013), which employs deep and domain-independent information extraction techniques, to extract concept maps from plain text documents. Though the goal of OntoCmap is not necessarily instructional, because it aims to learn domain ontologies through a large corpus of domain-related texts, it employs ideas and mechanisms that are very similar to those envisioned in this study. The OntoCmap technology is able to work with essay-length texts (e.g., expert text explanation) and to generate a variety of graph-based metrics along with traditional information retrieval measures such as TF and TF-IDF. Therefore, OntoCmap was an appropriate choice to generate concept maps and test the best metrics for identifying key concepts in a problem-solving task.

2.4. Metrics for key concept extraction

A concept map is a network of concepts—a graphical representation. Accordingly, analyzing a concept map relies on network analysis techniques that involve mathematical algorithms derived from graph theory (Rupp, Sweet, & Choi, 2010; Schvaneveldt, Durso, & Dearholt, 1989; Wasserman & Faust, 1994). Such analysis computes various graph-based metrics that characterize individual concept maps. Graph-based metrics include global-level and local-level measures (Wasserman & Faust, 1994). The global measures (e.g., centralization, size, density, clustering, and path length) describe the concept map as a whole, while local measures (e.g., betweenness centrality, degree) weigh each concept within a concept map. Based on the interest of graph-based measures in knowledge analysis (XXX, 2015; (Pirnay-Dummer, et al, 2009) and ontology learning (Mihalcea & Tarau, 2004), the current study assumes that graph-based metrics are reliable and viable measures through which key concepts can be identified.

There also statistical methods that can be used to extract key concepts from a text. For example, AutoTutor uses latent semantic analysis (LSA) to build a semantic space for computer literacy by analyzing two textbooks, 30 articles, and curricular materials (Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & Group, 1999). However, LSA is not well suited for short paragraphs or small amounts of text (XXX, 2013; Lintean et al., 2012). Besides LSA, Term Frequency (TF) and Term Frequency–Inverse Document Frequency (TF-IDF) are popular measures for identifying important words that occur more than once in a sample (Lintean et al., 2012).

This study compared computer-generated key concepts (according to candidate graph-based filtering metrics) to human expert-judged key concepts to investigate the most reliable methods for concept extraction. In addition, we compared these metrics to two information retrieval baselines, TF and TF-IDF.

3**. Research Methods**

3.1. Research questions

Based on the aforementioned theoretical and methodological assumptions, the following research questions were explored in this study:

- RQ1: How can we create key concepts for a complex problem-solving task? Do experts build a common problem space (i.e., common ground on a problem situation)?
- RQ2: To what extent do automatic filtering methods accurately extract key concepts? Do graph-based metrics outperform traditional information retrieval metrics (e.g., TF)?
- RQ3: Is it plausible and feasible to extract key concepts from a collective corpus of textual explanations by many experts instead of a singular ideal/expected textual explanation refined by experts?

This study conducted a sequence of analyses to answer the three research questions. First, we assumed that human experts in a particular domain are the best guides in obtaining an expert model. Accordingly, a carefully selected group of human experts (described in *3.2. Participants*) composed essay-length explanations of a given complex problem (see details in *3.3. Problem-solving situation*). Instead of paragraph-length responses, we used essay-length explanations (i.e., about 350-400 words) to locate key concepts on the grounds that there would be sufficient words to express the complexity of a problem.

Individual experts then participated in the Delphi process (Goodman, 1987; Hsu & Sandford, 2007) through which experts in the same domain can negotiate their expertise and build an ideal textual explanation of a complex problem. The Delphi method enabled the experts to agree on a set of key concepts that would be used as a benchmark. Moreover, the Delphi process served to investigate whether these participating experts exhibited different knowledge representations of the same problem (described in *3.4. Delphi method: A robust approach to building an expert model*).

Second, this study employed the OntoCmap tool to extract concepts from experts' written explanations and rank them based either on graph-based metrics or on traditional methods such as TF. OntoCmap generated diverse metrics that assigned a score to each concept extracted from a textual explanation. Filtered concepts based on the various metrics were then compared to the benchmark concepts selected by the human experts.

Third, this study investigated whether the automatic filtering methods were able to locate key concepts from a corpus of explanations as accurately as the ideal/expected explanation constructed by human experts in advance. Accordingly, this study compared the accuracy of multiple metrics obtained from a singular ideal/expected explanation and a corpus of texts written by seven experts, each of which was seemingly dissimilar to the others in word volume and embedded concepts and propositions.

3.2. Participants

Seven professors teaching at six major universities in the United States participated in a Delphi survey to obtain an expert model. Considering this study's problem situation, technology integration in a school curriculum, the panel members were selected based on pre-set criteria: (a) professors in Instructional Technology or related fields; (b) professors teaching a course entitled Instructional Design or Technology Integration in Learning; (c) professors who research technology-integration in classroom learning; and (d) professors whose doctorates were received at least three years prior to the study.

3.3. Problem-solving situation

This study used one of the problem-solving tasks from a course at a university in the southern United States. The course was designed to teach pre-service teachers how to integrate technology into teaching and learning. Students learned diverse technology tools and designed technology-supported/enhanced lesson activities to meet their particular needs. As part of the classroom activities, students composed written responses to complex problem situations to increase their awareness of the complex nature of educational problems and to activate their prior knowledge.

The problem situation used in the current study is based on the following scenario. An urban middle school attempted to integrate tablet PCs into in-classroom teaching and learning; however, this initiative had very little effect on the instructional practice in the classrooms (see Appendix A). The experts were asked to describe, in the form of written explanation, what concepts, issues, factors, and variables were likely to have contributed to the failure of the project (besides students), working under the assumption that they were evaluating this media implementation project. The purpose of their participation was to provide a reference against which student responses would be evaluated.

3.4. Delphi method: A robust approach to building an expert model

Using Delphi survey procedures (see Table 1), the seven experts co-constructed an ideal/expected explanation of the problem situation (see Appendix B). Each expert wrote his/her own explanation of the problem in the first round. Next, all of the statements and associated concepts in the seven explanations were qualitatively analyzed and consolidated for use in the second round. When aggregating the responses, the research team tried to retain all unique views while removing any redundancies. To represent multiple similar statements, a single statement was constructed based on the wording of the various statements.

— Insert Table 1 here —

In the second round, the experts were provided a full list of responses and asked to comment on the other experts' ideas and rank them. The experts' responses resulted in a list of key statements and concepts. Drawing on the selected key statements, the research team made a draft of an ideal/expected explanation in which key concepts were highlighted. In the final round, the experts again reviewed the ranking results and revised the draft version of an ideal/expected explanation, where necessary. There was little need for change. Consequently, they agreed on an ideal/expected explanation that embedded 23 key concepts.

3.5. OntoCmap: Automatic extraction of ranked concepts from textual explanations

The selected OntoCmap processed two types of textual data: (a) the ideal/expected explanation obtained through the Delphi process and (b) the corpus of seven expert explanations. OntoCmap performed several steps to identify key concepts. First, the tool used deep NLP to analyze textual information. Once all of the sentences in the corpus were processed, OntoCmap identified candidate concept-to-concept relations using syntactic patterns based on a dependency grammar representation (e.g., “technology implementations (concept 1)”-“begins with (predicate)”-“instructional need (concept 2)”; see Figure 2).

— Insert Figure 2 here —

These extracted individual concepts and relations were then combined into a concept map. The hypothesis of OntoCmaps is that the importance of concepts is directly reflected in the properties of those concepts and their relations (e.g., connectedness and degree). These measures were used to weigh the candidate concepts in the concept map (i.e., a graph). After the extraction of concept maps, OntoCmap assigned several scores to the obtained concepts using graph-based metrics as well as traditional information retrieval metrics. These metrics were then used to filter key concepts. Lastly, OntoCmap provided measures that indicated the performance of OntoCmap-generated key concepts against the benchmark. The automatically generated metrics are detailed in the next section.

3.6. Relevance metrics

Given a concept map where nodes represent potential concepts and edges are relations between those concepts, computing several measures based on graph properties is possible. In particular, previous metric comparison studies have shown an interest in Degree and Reachability (XXX, 2014) while other studies (XXX, 2013; XXX, 2011b) have shown that Degree and Betweenness yielded the most precise results. Accordingly, the current study focused on these three graph-based measures: Betweenness, Degree, and Reachability (for more details, see XXX, 2011b):

- Betweenness quantifies the number of times a node acts as a bridge along the shortest path between two other nodes (Brandes, 2001). Betweenness centrality is based on the assertion that a concept can exert control over the interaction between other pairs of concepts in a network (Anthonisse, 1971; Freeman, 1977).
- Degree represents the number of edges incident upon a node. Thus, the more a concept is connected to other concepts, the more it is considered important for the problem or situation.
- Reachability is calculated as the sum of the shortest distances between the candidate concept and any other reachable concept in the graph.
- For comparison purposes, this study also used two well-known metrics in information retrieval to assign a score to candidate concepts: TF and TF-IDF (Salton & Buckley, 1988).

Each of the selected metrics produces a list of potential concepts ranked based on their metric value in descending order. In order to determine key concepts from full-ordered lists, a threshold must be set (XXX, 2012). The mean threshold considers any concept value greater than or equal to the average value of the metric. We can also set a cut-off point using a particular percentage (e.g., over 25%) or number (Top-K). In other words, a given number of concepts are selected from the ranked lists. This study selected the first ranks threshold with *k*-20 because of the size of the benchmark (i.e., 23 key concepts) defined using the Delphi method.

3.7. Evaluation measures

Each metric derived key concepts from textual data. To evaluate these sets of key concepts against the benchmark, three standard types of evaluation metrics were used (Manning et al., 2008; Powers, 2011):

- Precision = items the metrics identified correctly / total number of items generated by the metrics

- Recall = items the metrics identified correctly / total number of relevant items (which the metrics should have identified)
- F-Measure = 2 * ((precision * recall) / (precision + recall))

4. Results

4.1. Expertise divergence, fast learning, and convergence

The seven experts built an expert model via the Delphi procedure. The initial round clearly revealed that individual experts represented various understandings of the same problem situation. As Table 2 describes, in contrast to the assumption that "experts would exhibit clearly recognizable patterns in their problem conceptualization, although experts did develop a variety of problem responses" (Spector, 2008, p. 31), the experts demonstrated diverse initial explanations to the same problem situation according to their prior experiences and individual interests. According to the experts, the number of issues ranged from 2 to 7(see Table 3). For example, experts E4 and E5 only included three and two issues, respectively, while extending them with sub-issues and supporting detailed evidence. Accordingly, these experts' explanations featured varying key concepts.

— Insert Table 2 here —

— Insert Table 3 here —

In the second round, the experts were provided a full list of statements (e.g., "Teachers were uncomfortable with their new roles"; "Teachers were simply teaching as they were taught") and concepts and asked to comments on other experts' ideas. The experts came across new ideas with which they extended or negotiated their conceptualization of the problem. After quickly learning about a diverse set of alternative opinions, these experts similarly ranked the issues from most to least important. Consequently, the experts agreed on six key issues (i.e., professional development, design issues, lack of supportive environment, empowerment, training, and mentoring), containing 23 key concepts (see Table 4).

— Insert Table 4 here —

4.2. Comparisons of relevance metrics

The OntoCmap tool generated expert models that were determined by various relevance metrics. These computer-generated expert models were compared to the benchmark created by human experts through the Delphi process. Table 5 summarizes the results of the comparisons. In the OntoCmap comparisons, the TF-IDF metric outperformed other metrics in precision, recall, and F-measure for both the ideal/expected explanation (Precision = 42.11, Recall = 15.38, and F-measure = 22.54) and the corpus of seven experts (Precision = 47.37, Recall = 17.31, and F-measure = 25.35). These results indicate that traditional standard measures performed better on an essay-length explanation. However, the values of recall were lower than 18% in both cases, meaning that less than 18% of key concepts were correctly identified. In pragmatic terms for teaching and learning, these low accuracy levels are problematic for precise assessment and targeted feedback.

— Insert Table 5 here —

For this reason, we conducted a qualitative inspection to identify the extent to which computer-generated key concepts matched the benchmark (see Table 6). Based on this visual inspection (Qualitative comparison), the contextual meaning of a word was also considered. For example, “decision-making” and “decision” were considered semantically equivalent. These supervised comparisons produced opposite results. Betweenness and Reachability were highest in all measures for the single data corpus (Precision = 55.00, Recall = 47.83, and F-measure =

51.16 for both metrics). Betweenness performance was highly tied to TF-IDF for the corpus of seven experts (Precision = 40.00, Recall = 34.78, and F-measure = 37.21).

— Insert Table 6 here —

4.3. The ideal/expected self-explanation (single data corpus) vs. the corpus of seven expert explanations (the corpus of seven experts)

As shown in Table 5, the single data corpus produced slightly higher average performance values than the corpus of seven experts when using metrics calculated by OntoCmap (Precision = 35.79, Recall = 13.08, and F-measure = 19.15; Precision = 34.74, Recall = 12.69, and F-measure = 18.59, respectively). However, the highest similarity measures were identified using TF-IDF for the corpus of seven experts (Precision = 47.37, Recall = 17.31, and F-measure = 25.35).

When using qualitative comparisons, the single data corpus clearly outperformed the corpus of seven experts on average (Precision = 52.00, Recall = 45.22, and F-measure = 48.37; Precision = 34.00, Recall = 29.57, and F-measure = 31.63, respectively). In addition, Betweenness and Reachability on the single data corpus produced the highest scores (Precision = 55.00, Recall = 47.83, and F-measure = 51.16).

Overall, the well-written "ideal" single explanation outperformed the corpus of seven experts in both OntoCmap-generated comparisons and qualitative comparisons. Performance measures suggested three most reliable filtering metrics: two graph-based metrics (i.e., Betweenness and Reachability) and one typical information retrieval metric (i.e., TF-IDF). Amongst these three metrics, Reachability performed consistently higher on the single data corpus.

As far as the recall measure (i.e., the ability to detect all key concepts) was concerned, the average difference between the single data corpus and the corpus of seven experts was 0.38 (3%) and 15.65 (16%) in the OntoCmap comparisons and qualitative comparisons, respectively. These findings imply that using a single ideal explanation is “ideal,” and that the use of a corpus of multiple explanations needs more testing to demonstrate better precision.

5. Discussion

5.1. Findings

The Delphi process confirmed the argument that even experts in the same domain have varying understandings of a particular problem situation even while sharing some views of the problem. The levels of variation could differ according to the nature of a problem (i.e., complexity and ill-structuredness).

The selected OntoCmap tool computed automatic filtering metrics that demonstrated about 18% detection accuracy in filtering correct key concepts. When reviewing semantic overlapping by visual inspection, the recall value increased to about 50%, especially when using graph-based metrics such as Betweenness and Reachability. The results show that automatic filtering metrics could be used to build an initial expert model for a particular problem in the context of instruction. However, detection accuracy needs improvement.

Interestingly, there was only a 3% difference in the OntoCmap recall measures between the single data corpus and the corpus of seven experts, though this difference increased to 16% after the qualitative comparisons (i.e., visual inspection by the research team). The seven experts in their initial text explanations represented quite dissimilar understandings of the problem situation. However, the corpus aggregating all of their explanations built up a somewhat connected semantic space. The filtering metrics presumably were able to extract an expert model that was close to an ideal expert model when drawing from a corpus of explanations by many experienced experts.

5.2. Implications

Automatic modeling of human cognition via NLP methods has been studied primarily in connection with ITSs (e.g., using textual information from conversations or paragraph-length explanations) (D'Mello et al., 2010; Nye et al., 2014; Rus et al., 2013; Sottilare, Graesser, Hu, & Goldberg, 2014). The current study uniquely applied automatic modeling methods to essay-length explanations (i.e., at least 350 words). We assumed that deep NLP and structural analysis of knowledge representations (i.e., graph-based modeling) would be useful in building more accurate expert models.

The findings of this study demonstrate that one can extract an initial expert model (i.e., key concepts) for a particular problem via automatic methods. In particular, graph-based metrics such as Betweenness and Reachability best detected key concepts from the single data corpus. Moreover, instead of winding through labor and time-intensive methods such as the Delphi process to construct a single data corpus, automatic modeling methods can extract an expert model using a collection of explanations by a group of experts. Yet automatic and precise formative assessment and feedback still requires the participation of human experts, who create an ideal/expected explanation that more fully accounts for a complex problem space. In addition, human experts can set a filtering threshold that befits the nature of the problem.

The proposed methods can be applied to an expert-modeling module in diverse technology-enhanced learning support systems (i.e., in-classroom support systems, learning games and simulations, ITSs, and data visualization and learning analytics in Learning Management Systems) in which expert models such as key concepts are used to evaluate student understanding. For example, as a potential application, we can imagine an in-classroom support system similar to Flicker but for problem-centered learning. Using the system, an earth science teacher could create an authentic problem situation regarding the frequent wildfires in California in 2015. Concerning the climate change problem, the teacher could write an ideal explanation for the following question: "Which factors might contribute to the frequent wildfires in CA in 2015?" The embedded methods could automatically analyze the teacher's explanation and build an expert model. Students would explore a simulated real-world problem on their tablets to answer the question. The adaptive system could gather and analyze the students' explanations, using the expert model as a reference. Finally, the teacher could measure class-level understanding or individual learning progress. The system could then suggest appropriate feedback strategies tailored to a selected individual or the whole class by pointing out specific concepts or propositions.

5.3. Suggestions

In spite of the potential benefits of this study, admittedly, correctly extracting concepts and their relations from textual data and building an appropriate concept map (i.e., a graph) from a corpus of limited size remain difficult tasks. Compared to traditional standard measures such as TF, graph-based measures are very sensitive to the structure of a concept network. For example, XXX (2012, 2014) found that different methods for distilling concepts and relations produced quite dissimilar concept maps. These findings warrant further studies to explore more feasible and reliable methods that accurately extract semantically meaningful concepts and relations from textual data. Below are several suggestions that can guide future research.

Future studies need to address the best NLP algorithms to identify semantic components (i.e., concepts and relations) for the purpose of expert modeling and assessing student models.

Tuning some of the extraction patterns on small amounts of text could potentially improve performance. For example, a single textual explanation can create different concept maps depending on concept filtering (e.g., are there two concepts in the following examples?) based on noun-compounds (e.g., "knowledge analysis") or words-dependencies (e.g., prepositional phrases attached to nouns; "success of active learning"). Determining semantic similarity is another issue (e.g., Does "analysis" overlap with "knowledge analysis" in the problem context?).

There could be several competing extraction patterns depending on the way concepts are defined and their paired relations in a textual explanation. Accordingly, future studies need to elaborate extraction patterns and compare them with the relevance metrics and evaluation measures used in the current study. Clear guidance about the number of words necessary for an expert to explain the nature of a target problem is important for several reasons: (a) the higher the complexity of the problem, the more words are needed to explain its nature; (b) the number of words is directly associated with the number of concepts that comprise a concept map; (c) the structural complexity of a concept map relates to or determines the values of most graph-based metrics; and (d) concerning assessment implementation, the number of words influences the time needed to complete a textual response to a problem. For example, based on previous studies (XXX, 2012, 2014), the current study assumed that 350-450 words would be enough to describe a complex problem situation.

Future research could further investigate the potential use of multiple explanations as a data set. Though the current study demonstrated that an ideal/expected explanation was more precise in key concept extraction than a collection of explanations, we can postulate that a larger corpus of expert explanations could yield an equivalent level of precision. Moreover, newly tuned extraction patterns could produce different findings. Provided that there were more accurate and reliable ways to extract key concepts from a large collection of explanations, one could even use a large corpus of student explanations to identify key concepts that occur in a single ideal explanation. We can assume that key concepts would far more often exist across a group of student explanations, yielding a collective concept map that could subsequently be filtered by graph-based metrics. Building on either a collection of expert explanations or a much larger number of student responses, advanced and validated automatic expert modeling methods could make the labor-intensive process of writing an "ideal" single explanation unnecessary.

Lastly, the filtering metrics that were applied to building an expert model can be used to extract student models for a particular problem. The performance measures (e.g., precision, recall, and F-measure) could be used as similarity measures (i.e., between an expert model and a student model) that determine levels of student understanding (see other similarity measures; XXX, 2015). These computerized assessment and feedback mechanisms can advance the collaborative reasoning process in a learning system (Graesser, Person, & Magliano, 1995): (a) posing a problem, (b) initial student explanation, (c) brief evaluation, (d) scaffolding (i.e., hint, deep questioning, and mini lectures) based on similarity analysis, and (e) student modification of his/her initial answer.

5.4. Limitations

This study's inquiries included whether deep NLP and graph-based metrics are able to construct an expert model using a corpus of expert textual data instead of an ideal/expected explanation. The question was prompted by Lintean et al. (2012), which attempted to build an expert model using a large corpus of paragraph-length student explanations without any input from an expert. The current study proved that a collection of expert texts could, to some degree, construct an expert model similar to an ideal/expected explanation using several essay-length explanations. Accordingly, whether an expert model can be built without careful revision by experts themselves remains debatable.

References

Alexander, P. A. (2004). A model of domain learning: Reinterpreting expertise as a multidimensional, multistage process. In *Motivation, emotion, and cognition: Integrative perspectives on intellectual functioning and development* (pp. 273–298). Mahwah, NJ: Lawrence Erlbaum Associates.

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind?: Modeling students' reading comprehension skills with natural language processing techniques. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 246–254). New York, NY, USA: ACM. http://doi.org/10.1145/2723576.2723617

Anderson, J. R. (1980). Cognitive psychology and it's implications. *San Francisco: Freman*, *119*.

André A. Rupp, Shauna J. Sweet, & Younyoung Choi. (2010). Modeling learning trajectories with epistemic network analysis: A simulation-based investigation of a novel analytic method for epistemic games.

Anthonisse, J. M. (1971). The rush in a graph. *Mathematische Centrum, Amsterdam*.

Axelrod, R. (2015). *Structure of decision: The cognitive maps of political elites*. Princeton university press.

Baker, R. S. J. d, & Siemens, G. (n.d.). Educational data mining and learning analytics. Retrieved from www.columbia.edu/~rsb2162/BakerSiemensHandbook2013.pdf

Brandes, U. (2001). A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, *25*(2), 163–177.

Brown, D. (1997). *An introduction to object-oriented analysis: objects in plain English*. Wiley New York.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*, 55–67.

Chi, M. T. H. (2006). Laboratory methods for assessing experts' and novices' knowledge. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511816796.010

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*(3), 439–477. http://doi.org/10.1016/0364-0213(94)90016-7

Clariana, R. B., & Wallace, P. (2009). A comparison of pair-wise, list-wise, and clustering approaches for eliciting structural knowledge in information systems courses. *International Journal of Instructional Media*, *36*(3), 287–302. http://doi.org/Article

Clariana, R. B., Wolfe, M. B., & Kim, K. (2014). The influence of narrative and expository lesson text structures on knowledge structures: alternate measures of knowledge structure. *Educational Technology Research and Development*, *62*(5), 601–616. http://doi.org/10.1007/s11423-014-9348-3

Clariana, R., Wallace, P., & Godshalk, V. (2009). Deriving and measuring group knowledge structure from essays: The effects of anaphoric reference. *Educational Technology Research and Development*, *57*(6), 725–737. http://doi.org/10.1007/s11423-009-9115-z

Coronges, K. A., Stacy, A. W., & Valente, T. W. (2007). Structural comparison of cognitive associative networks in two populations. *Journal of Applied Social Psychology*, *37*(9), 2097–2129. http://doi.org/10.1111/j.1559-1816.2007.00253.x

Dascalu, M., Trausan-Matu, S., Dessus, P., & McNamara, D. S. (2015). Discourse cohesion: A signature of collaboration. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (pp. 350–354). New York, NY, USA: ACM. http://doi.org/10.1145/2723576.2723578

D'Mello, S., Hays, P., Williams, C., Cade, W., Brown, J., & Olney, A. (2010). Collaborative lecturing by human and computer tutors. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent Tutoring Systems* (pp. 178–187). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.proxy-remote.galib.uga.edu/chapter/10.1007/978-3-642-13437-1_18

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). Methods for studying the structure of expertise: psychometric approaches. In *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9780511816796

Feyzi-Behnagh, R., & Azevedo, R. (2012). The effectiveness of a pedagogical agent's immediate feedback on learners' metacognitive judgments during learning with MetaTutor. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent Tutoring Systems* (Vol. 7315, pp. 651–652). Springer Berlin / Heidelberg. Retrieved from http://www.springerlink.com/content/x58705l375254652/abstract/

Flavell, J. H. (1992). Cognitive development: Past, present, and future. *Developmental Psychology*, *28*(6), 998–1005.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2009). Classification of semantic relations between nominals. *Language Resources and Evaluation*, *43*(2), 105–121. http://doi.org/10.1007/s10579-009-9083-2

Glaser, R., Chi, M. T., & Farr, M. J. (1988). *The nature of expertise*. Lawrence Erlbaum Associates Hillsdale, NJ.

Gobet, F., & Wood, D. (1999). Expertise, models of learning and computer-based tutoring. *Computers & Education*, *33*(2–3), 189–207.

Goodman, C. M. (1987). The Delphi technique: a critique. *Journal of Advanced Nursing*, *12*(6), 729–734.

Gopal, T., Herron, S. S., Mohn, R. S., Hartsell, T., Jawor, J. M., & Blickenstaff, J. C. (2010). Effect of an interactive web-based instruction in the performance of undergraduate anatomy and physiology lab students. *Computers & Education*, *55*(2), 500–512.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, *9*(6), 495–522.

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & Group, T. R. (1999). AutoTutor: A simulation of a human tutor. *Cognitive Systems Research*, *1*(1), 35–51.

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical Assessment, Research & Evaluation*, *12*(10), 1–8.

Janssen, T. M. V. (2012). Montague semantics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2012). Retrieved from http://plato.stanford.edu/archives/win2012/entries/montague-semantics/

Johnson-Laird, P. N. (2013). Mental models and cognitive change. *Journal of Cognitive Psychology*, *25*(2), 131–138.

Jonassen, D. H. (2014). Assessing problem solving. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 269–288). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-3185-5_22

Jonassen, D. H., Beissner, K., & Yacci, M. (1993). Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge. *Structural Knowledge: Techniques for Representing, Conveying, and Acquiring Structural Knowledge.*

Jonnassen, D. H., & Philip, H. (1996). Mental models: knowledge in the head and knowledge in the world. In *Internaltion Conference on Learning Science* (pp. 433–438). Evanston, Illinois: International Society for the Learning Science.

Authors (2012). ---

Authors (2013). ---

Authors (2012). ---

Authors (2015). ---

Kouznetsov, A., & Zouaq, A. (2014). A comparison of graph-based and statistical metrics for learning domain keywords. In Y. S. Kim, B. H. Kang, & D. Richards (Eds.), *Knowledge Management and Acquisition for Smart Systems and Services* (pp. 260–268). Springer International Publishing.

Lintean, M., Rus, V., & Azevedo, R. (2012). Automatic detection of student mental models Based on natural language student input during metacognitive skill training. *International*

Journal of Artificial Intelligence in Education, *21*(3), 169–190. http://doi.org/10.3233/JAI-2012-022

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

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499–515. http://doi.org/10.3758/s13428-012-0258-1

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into texts. *Association for Computational Linguistics*.

Molloy, E. K., & Boud, D. (2014). Feedback models for learning, teaching and performance. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 413–424). Springer New York. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4614-3185-5_33

Montague, R. (1974). In*formal philosophy: Selected papers of Richard Montague; Ed. and with an Introduction by Richmond H. Thomason*. Yale University Press.

Narayanan, V. K. (2005). Causal mapping: An historical overview. *Causal Mapping for Research in Information Technology*, 1–19.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104). Prentice-Hall Englewood Cliffs, NJ.

Nkambou, R. (2010). Modeling the domain: An introduction to the expert module. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems* (pp. 15–32). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-14363-2_2

Novak, J. D., & Cañas, A. J. (2006). The origins of the concept mapping tool and the continuing evolution of the tool. *Information Visualization*, *5*(3), 175–184.

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427–469. http://doi.org/10.1007/s40593-014-0029-5

Pirnay-Dummer, P., & Ifenthaler, D. (2010). Automated knowledge visualization and assessment. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (pp. 77–115). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-5662-0_6

Pirnay-Dummer, P., Ifenthaler, D., & Spector, J. (2009). Highly integrated model assessment technology and tools. *Educational Technology Research and Development*. Retrieved from http://dx.doi.org/10.1007/s11423-009-9119-8

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Pretz, J. E., Naples, A. J., & Sternberg, R. J. (2003). Recognizing, defining, and representing problems. *The Psychology of Problem Solving*, *30*(3).

Rus, V., D’Mello, S., Hu, X., & Graesser, A. (2013). Recent advances in conversational intelligent tutoring systems. *AI Magazine*, *34*(3), 42–54. http://doi.org/10.1609/aimag.v34i3.2485

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

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. *The Psychology of Learning and Motivation*, *24*, 249–284.

Seel, N. M. (2003). Model-centered learning and instruction. *Technology, Instruction, Congnition and Learning*, *1*, 59–85.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning.

Shermis, M. D. (2010). Automated essay scoring in a high stakes testing environment. In V. J. Shute & B. J. Becker (Eds.), *Innovative Assessment for the 21st Century* (pp. 167–185). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-6530-1_10

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 403–416). New York: The Guilford Press.

Shute, V. J., Masduki, I., Donmez, O., Dennen, V. P., Kim, Y.-J., Jeong, A. C., & Wang, C.-Y. (2010). Modeling, assessing, and supporting key competencies within game environments. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-Based Diagnostics and Systematic Analysis of Knowledge* (pp. 281–309). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-5662-0_15

Sottilare, R., Graesser, A., Hu, X., & Goldberg, B. (2014). *Design recommendations for intelligent tutoring systems: Volume 2: Instructional Management*. Army Research Laboratory.

Spector, J. M. (2008). Complex domain learning. In *Handbook on Information Technologies for Education and Training* (pp. 261–275).

Spector, J. M. (2010). Mental representations and their analysis: An epistemological perspective. In D. Ifenthaler, P. Pirnay-Dummer, & N. M. Seel (Eds.), *Computer-Based Diagnostics* *and Systematic Analysis of Knowledge* (pp. 27–40). Springer US. Retrieved from http://link.springer.com/chapter/10.1007/978-1-4419-5662-0_3

Spector, J. M., & Koszalka, T. A. (2004). The DEEP methodology for assessing learning in complex domains (Final report to the National Science Foundation Evaluative Research and Evaluation Capacity Building). *Syracuse, NY: Syracuse University*.

Wasserman, S., & Faust, K. (1994). *Social network analysis: methods and applications*. Cambridge University Press.

Xu, F., & Krieger, H.-U. (2003). Integrating shallow and deep NLP for information extraction. In *RANLP 2003*.

Authors (2010). ---

Authors (2011a). ---

Authors (2011b). ---

Authors (2012). ---

Authors (2013). ---

Table 1

Delphi Procedure

Round	Activity
R1. Brainstorming	• Collect and consolidate all responses from experts
R2. Narrowing Down/ Ranking	• Send refined final version of consolidated lists, including statements and used concepts • Ask experts to add comments if they disagree with or have different opinions about a statement • Ask experts to rank key statements and concepts
R3. Refinement	• Send each panelist ranked statements and concepts summarized by the investigators • Ask for revision of judgments or specification of reasons for remaining outside the consensus

Note. XXX (2013, p. 960).

Table 2

Identification of Issues via Content Analysis

#[a]	Issues	E1[b] (N = 7)	E2 (N = 7)	E3 (N = 4)	E4 (N = 3)	E5 (N = 2)	E6 (N = 5)	E7 (N = 4)
1	Community of Practice		Community of Practice					
1	Student						Grade level	
1	Classroom as an ecosystem	Classroom as an ecosystem						
2	High-stakes Testing	state-mandated testing	high-stakes tests					
2	Mentoring		mentor					mentor
2	Empowerment & Engagement		Top-down decision making			Participatory design		
2	Barriers to Technology	Barriers to Technology		lack of time to transform lessons				
3	Design Issue	lack of definition for desired performance			clearly identified educational need	Meaning of artifacts; Primary generator		
3	Teacher Belief and Attitude	Teacher Belief			agreement that technology is the solution		attitude	
3	Need for Professional Development		Professional Development	professional development				professional development
4	Resisting Change	Resist innovation	uncomfortable with their new role	teachers are simply teaching			the length of time they had been teaching	
4	Lack of Supportive Environment	Environment			visible, sustainable commitment		Conservative cultural environment; incentives	supportive environment
4	Insufficient Training		training	Training			training	training

Note. a. The numbers in column #a indicate the number of agreements among panel members for the issues, b. "*E"* denotes an expert, and the following number in parenthesis indicates the number of issues the expert mentioned in his/her textual explanation.

Table 3

Identification of Key Terms via Content Analysis

Issue	Key Terms (frequency)
C1. Resisting Change	Innovation (4), New role (4), Habit (3)
C2. Lack of Supportive Environment	Environment (5), Incentive (3), Commitment (2), Change (1), Instructional practice (2), Culture (1), Performance (1)
C3. Insufficient Training	Integration (4), Pedagogical use (2), Training (3), Time (1)
C4. Design Issue	Integration (3), Assumption (4), Instructional need (2), Design space (2)
C5. Belief and Attitude	Belief (4), Interest (2), Need (2)
C6.Professional Development	Professional development (3), Affordance (3), Best practice (1), Pedagogy (4)
C7. High-Stakes Tests	Test (3)
C8. Mentoring	Mentor (4), Support (1), Motivation (2)
C9.Empowerment & Engagement	Buy into (2), Collaborators (2), Decision (3), Intervention (1), New media (1)
C10.Community of Practice	Community of practice (5)
C11. Students	Grade level (2), Game (1), Advantage (1)

Note. The experts chose two or three of the most important key terms associated with each key issue.

Table 4

Ranks of the Issues

RANK	Issue	Ave.	Min	Max	Range	Median
1	C6. Professional Development	3.2	1	7	6	3
2	C4. Design Issue	4.0	1	8	7	3
3	C2. Lack of Supportive Environment	4.2	1	8	7	4
3	C9. Empowerment & Engagement	4.3	2	8	6	3
5	C3. Insufficient Training	5.0	1	11	10	5
5	C8. Mentoring	5.0	4	8	4	4.5
7	C5. Belief and Attitude	6.5	2	11	9	7
8	C1. Resisting Change	6.5	2	9	7	6.5
9	C7. High-Stakes Tests	8.5	5	11	6	9.5
10	C10. Community of Practice	8.8	5	11	6	9
11	C11. Students	10.0	9	11	2	10

Note. The experts ranked the key issues in order of their contribution to the poor project results from 1 (most influential) to 11 (least influential).

Table 5

Comparison of Relevance Metrics

		OntoCmap Comparisons			Qualitative Comparisons			
		Precision	Recall	F-Measure	Matched N	Precision	Recall	F-Measure
Single	Betweenness	26.32	9.62	14.08	11.00	55.00	47.83	51.16
	Degree	31.58	11.54	16.90	10.00	50.00	43.48	46.51
	Reachability	36.84	13.46	19.72	11.00	55.00	47.83	51.16
	TFIDF	42.11	15.38	22.54	10.00	50.00	43.48	46.51
	TF	42.11	15.38	22.54	10.00	50.00	43.48	46.51
	Min	26.32	9.62	14.08	10.00	50.00	43.48	46.51
	Max	42.11	15.38	22.54	11.00	55.00	47.83	51.16
	Average (A)	35.79	13.08	19.15	10.40	52.00	45.22	48.37
Corpus	Betweenness	36.84	13.46	19.72	8.00	40.00	34.78	37.21
	Degree	31.58	11.54	16.90	6.00	30.00	26.09	27.91
	Reachability	21.05	7.69	11.27	6.00	30.00	26.09	27.91
	TFIDF	47.37	17.31	25.35	8.00	40.00	34.78	37.21
	TF	36.84	13.46	19.72	6.00	30.00	26.09	27.91
	Min	21.05	7.69	11.27	6.00	30.00	26.09	27.91
	Max	47.37	17.31	25.35	8.00	40.00	34.78	37.21
	Average (B)	34.74	12.69	18.59	6.80	34.00	29.57	31.63
Average Difference (A-B)		1.05	0.38	0.56	3.60	18.00	15.65	16.74

Table 6

Supervised Review of Semantic Overlapping of the Single Data Corpus

Gold Standards	Betweenness (N = 11)	Degree (N = 10)	Reachability (N = 11)	TFIDFDT (N = 10)	TFDT (N = 10)
affordance	affordances_of_tablets	affordances_of_tablets	affordances_of_tablets		
assumption	Assumptions				
attitude					
best_practice	Practices	practices		practices	practices
change				changing	changing
collaborator					
culture					
decision_making	decision	decision	decision		
design_space	design			design	design
environment		ongoing_supportive_environment	ongoing_supportive_environment	environment	environment
incentive	advantage	lack_of_incentives	lack_of_incentives		
instructional_need		instructional_need	instructional_need	instructional_need	instructional_need
instructional_practice			instructional_practices	instructional_practices	instructional_practices
integration					
lecture					
mentor	mentor			mentor	mentor
motivation					
outcome					
pedagogy	pedagogy	pedagogy	pedagogy		
professional_development			teacher_professional_development	professional_development	professional_development
support	technical_support	technical_support	support	support	support
teacher_belief	teacher_beliefs	teacher_beliefs	teacher_beliefs	beliefs	beliefs
use of technology	effective_use_of_new_technology	uses_of_technology	use_of_technology		

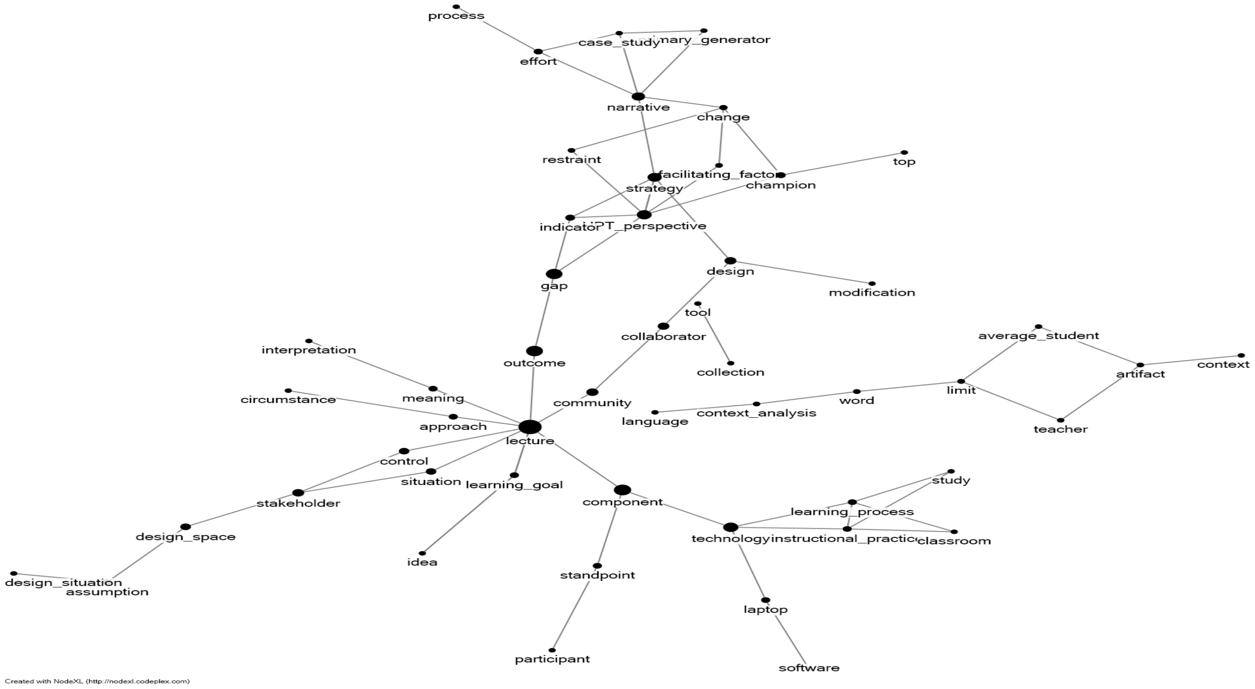

Figure 1. An expert's concept map

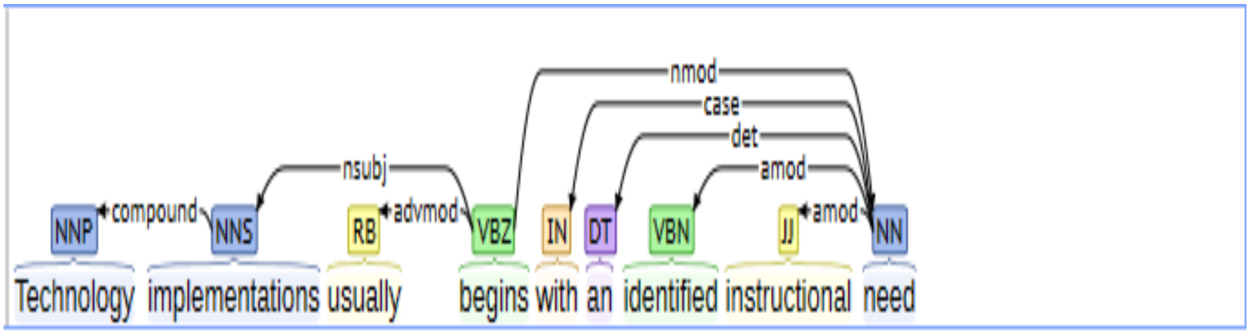

Figure 2. Sentence dependency representation, pattern example, and triple extraction [1]

[1] Visual dependency representation generated using http://nlp.stanford.edu:8080/corenlp/process

Appendix A

Case Study

Directions: Read the case study described below and then prepare a response to the questions below (**written response with at least 350 words is required for each question**):

Assume that you have been involved in evaluating a media implementation project in an urban inner middle school. At the beginning of the school year, all of the students assigned to four subject area teachers (math, language arts, social studies and science) in the seventh grade at the middle school were given tablet PCs (laptop computers also equipped with a stylus/pen and a touchscreen that can be written upon) and were also given wireless internet access at home and at school for a entire year.

The students took the tablet PCs home every evening and brought them to class every day. The teachers were also provided with tablet PCs 24/7 (24 hours a day, every day of the week) for the entire year. The teachers and students were trained on how to use the tablet PCs. Moreover, all of the curriculum materials (textbooks, workbooks, student study guides, teacher curriculum guides, some activities, tests, etc.) were installed on the tablet PCs or were accessible through the tablet PCs.

Your job as one of the evaluators for the project was to examine how this innovation (providing teachers and students with tablet PCs 24/7) changed the way instruction was presented in the classrooms of the four teachers. Results indicated that the innovation had very little effect on the manner in which instruction took place in the teachers' classrooms.

1. Based on what you have learned about the use of technology in education, describe what concepts, issues, factors, and variables are likely to have contributed to the fact that the introduction of the tablet PCs had very little effect on the instructional practices that were employed in the classes.

2. Describe the strategies that could have been employed to help mitigate the factors that you think contributed to the minimal effect the tablet PCs had on instructional practices. When you answering this question, use the concepts, factors, and variables you described in the question 1 or add other assumptions and information that would be required to solve this problem.

Appendix B

An ideal/expected explanation

Technology implementations usually begin with an identified instructional need. Instructional need was likely not fully identified due to insufficient study of how instructional practices in the classroom were being conducted already without the technology. One big issue is defining what a successful integration or change in instructional practice actually is. While teachers in the situation may have felt that they knew this already, the assumptions inherent in a design situation need to be articulated and checked if the assumptions are not to distort the design space by which instructional practices are manipulated. Teachers didn't have enough professional development using the technology in classroom teaching and learning, on ways to integrate use into their teaching, and best practices with regard to effective educational use. Teacher professional development that discusses not just technical know-how but also pedagogy could help teachers realize how to do things differently that takes full advantage of the affordances of the tablets. Training as a professional development should be extensive including teacher beliefs and attitude. Teacher beliefs play a role in adopting new practices and changing their instructional practice. Teachers may not believe that students learn with laptops, and thus do not use laptops in their instruction. The only support teachers had during implementation was technical support; Teachers lacked a mentor who could assist them as instructional issues arose throughout the year. Mentoring on additional and advanced uses of the technology in the classroom is critical for teachers to increase their skills and maintain their motivation in utilizing the technology. In addition, mentor could help teachers to maintain the belief that these efforts will have positive results. There are concerns that the environment does not support change. An ongoing supportive environment where teachers initially learn how to use the technology, how to use the technology with their content, and how to continue to develop their expertise in the technology and incorporating it to the classroom is critical.

Environment could include a culture that does not support the desired performance.

For example, the lack of incentives to make effective use of a new technology could also contribute to lack of use. The intervention seems to have been applied to this community rather than involving teachers from the beginning as collaborators in its design and modification. Teachers were not involved in the decision to implement the new media; thus, they did not fully "buy into" the plan.