Dialogue and Discourse 3(2) (2012) 75–99 doi: 0.5087/dad.2012.204

Question Generation from Concept Maps

Andrew M. Olney AOLNEY@MEMPHIS.EDU
Arthur C. Graesser A-GRAESSER@MEMPHIS.EDU
Institute for Intelligent Systems
University of Memphis
365 Innovation Drive Memphis, TN 38152

Natalie K. Person PERSON@RHODES.EDU
Department of Psychology
Rhodes College
2000 N. Parkway, Memphis, TN 38112-1690

Editors: Paul Piwek and Kristy Elizabeth Boyer

Abstract

In this paper we present a question generation approach suitable for tutorial dialogues. The approach is based on previous psychological theories that hypothesize questions are generated from a knowledge representation modeled as a concept map. Our model semi-automatically extracts concept maps from a textbook and uses them to generate questions. The purpose of the study is to generate and evaluate pedagogically-appropriate questions at varying levels of specificity across one or more sentences. The evaluation metrics include scales from the Question Generation Shared Task and Evaluation Challenge and a new scale specific to the pedagogical nature of questions in tutoring

1. Introduction

A large body of research exists on question-related phenomena. Much of this research derives from the tradition of question answering rather than question asking. Whereas question answering has received extensive attention in computer science for several decades, with increasing interest over the last 10 years (Voorhees and Dang 2005, Winograd 1972), question asking has received attention primarily in educational/psychological circles (Beck et al. 1997, Bransford et al. 1991, Brown 1988, Collins 1988, Dillon 1988, Edelson et al. 1999, Palinscar and Brown 1984, Piaget 1952, Pressley and Forrest-Pressley 1985, Scardamalia and Bereiter 1985, Schank 1999, Zimmerman 1989) until the recent surge of interest in computational accounts (Rus and Graesser 2009).

The distinction between question answering and question asking in some ways parallels the distinction between natural language understanding and natural language generation. Natural language understanding takes a piece of text and maps it to one of many possible representations, while natural language generation maps from many possible representations to one piece of text (Dale et al. 1998). Question answering, after all, requires some understanding of what the question is about in order for the correct answer to be recognized when found. Likewise question asking seems logically tied to the idea of selection, since only one question can be asked at a given moment.

Though the above description appeals to representations, the use of representations in question answering has been largely avoided in state of the art systems (Ferrucci et al. 2010, Moldovan et al. 2007). Given the success of representation-free approaches to question answering, a pertinent question is whether the same can be true for question generation (Yao 2010). The simplest instantiation of representation-free approach would be to syntactically transform a sentence into a question using operations like *wh-fronting* and *wh-inversion*, e.g. "John will eat sushi" becomes "What will John

 Submitted 3/11; Accepted 1/12; Published online 3/12

eat?" And indeed, this basic approach has been used extensively in computational models of question generation (Ali et al. 2010, Heilman and Smith 2009, 2010, Kalady et al. 2010, Varga and Ha 2010, Wyse and Piwek 2009). These approaches are knowledge-poor in that they do not take into account properties of words other than their syntactic function in a sentence.

One significant problem with knowledge poor approaches is determining the *question type*. In the example above, it is necessary to determine if "sushi" is a person, place, or thing and use "Who," "Where," or "What," respectively. So it appears that some degree of knowledge representation is required to generate questions with this higher degree of precision. The literature on question generation has approached this problem in primarily two different ways. The first is by using named entity recognizers to classify phrases into useful categories like place, person, organization, quantity, etc. (Ali et al. 2010, Kalady et al. 2010, Mannem et al. 2010, Varga and Ha 2010, Yao and Zhang 2010). The second approach to determining question type is semantic role labeling (Chen et al. 2009, Mannem et al. 2010, Pal et al. 2010, Sag and Flickinger 2008, Yao and Zhang 2010). Semantic role labelers assign argument structures to parse trees, specifying the subject, patient, instrument, and other roles associated with predicates (Palmer et al. 2005). Role labels thus make it easier to deal with phenomena like passivization which invert the order of arguments. Role labels also provide useful adjunct roles specifying causal, manner, temporal, locative, and other relationships (Carreras and Màrquez 2004). Previous work on question generation has demonstrated that named entities and adjunct roles may both be mapped to question type categories in a straightforward way. For example, the sentence

> "<Person>Charles Darwin</Person> was impressed enough with earthworms that he devoted years-and an entire book-to their study."

may be transformed into the question

> "Who was impressed enough with earthworms that he devoted years-and an entire book-to their study?"

by mapping the Person named entity "Charles Darwin" to the question type "Who." Likewise the sentence

> "<Am-Cau>Because fermentation does not require oxygen</Am-Cau>, fermentation is said to be anaerobic."

may be transformed into the question

> "Why is fermentation said to be anaerobic?"

by mapping the causal adjunct Am-Cau to the question type "Why."

Clearly semantic role labelers and named entity recognizers bring significant value to the question generation process by adding knowledge. One might speculate that if a little knowledge is good, perhaps a lot of knowledge is even better. But what sort of knowledge?

Recent work by Chen and colleagues has explored the role that knowledge may play in generating questions (Chen et al. 2009, Chen 2009, Mostow and Chen 2009). Rather than generating questions one sentence at a time, their system builds a situation model of the text and then generates questions from that model. In an early version of the system, which works exclusively with narrative text, Chen et al. (2009) employ the clever trick of maintaining a situation model only of characters' mental states. Because mental states of characters are usually fundamental in explaining the evolution of a narrative, questions generated from the model appear to satisfy a nontrivial challenge for question generation: generating *important* questions that span more than one sentence (Vanderwende 2008). Although the early version of the system only generates simple yes/no questions, later versions can generate "Why" and "How" questions for both narrative and informational text. An example of a mental state "Why" question for narrative text is give in Figure 1.

> Once upon a time a town mouse, on a trip to the country, met a country mouse. They soon became friends. The country mouse took his new friend into the meadows and forest. To thank his friend for the lovely time, he invited the country mouse to visit him in the town. And when the country mouse saw the cheese, cake, honey, jam and other goodies at the house, he was pleasantly surprised.
>
> Right now the question I'm thinking about is, why was the country mouse surprised?

Figure 1: An example passage and question from Mostow and Chen (2009)

All versions of the system use parsing and semantic role labeling to extract argument structures, with particular emphasis on modal verbs, e.g. "believe," "fear," etc, which are then mapped to the situation model, a semantic network. The mapping process involves a semantic decomposition step, in which modal verbs like "pretend" are represented as "X is not present in reality and person P1's belief, but P1 wants person P2 to believe it" (Chen 2009). Thus the network supports inference based on lexical knowledge, but it does not encode general world knowledge[1]. Questions are generated from this knowledge structure by filling templates (Chen et al. 2009, Mostow and Chen 2009):

- Why/How did <character> <verb> <complement>?
- What happens <temporal-expression>?
- Why was/were <character> <past-participle>?
- Why <auxiliary-verb> <x>?

The work of Chen et al. is much more closely aligned with psychological theories of question generation than work that generates questions from single sentences (Ali et al. 2010, Heilman and Smith 2009, 2010, Kalady et al. 2010, Mannem et al. 2010, Pal et al. 2010, Sag and Flickinger 2008, Varga and Ha 2010, Wyse and Piwek 2009, Yao and Zhang 2010) in at least two ways. By attempting to *comprehend* the text before asking questions, the work of Chen et al. acknowledges that question asking and comprehension are inextricably linked (Collins and Gentner 1980, Graesser and Person 1994, Hilton 1990, Olson et al. 1985). Additionally, Chen et al.'s work underscores the pedagogical nature of questions, namely that question asking can involve a tutor on behalf of a student as well as the student alone.

The primary goal of the research reported here is to continue the progress made by Chen et al. in connecting the computational work on question generation with previous research on question asking in the psychological literature. This study attempts to close the psychological/computational gap in two ways. First, we review work from the psychological literature that is relevant to a computational account of question generation. Secondly, we present a model of tutorial question generation derived from previous psychological models of question asking. Thirdly, we present an evaluation of this model using a version of the Question Generation Shared Task and Evaluation Challenge metrics (Rus et al. 2010a; Rus et al., this volume) augmented for the pedagogical nature of this task.

2. Psychological Framework

In tutorial contexts, question generation by both human instructors (tutors) and students has been observed by Graesser, Person, and colleagues (Graesser and Person 1994, Graesser et al. 1995, Person

1. World knowledge is often included in situation models (McNamara and Magliano 2009).

Table 1: Graesser, Person, and Huber (1992)'s Question Categories

Question category	Abstract specification
1. Verification	Is X true or false? Did an event occur?
2. Disjunctive	Is X, Y, or Z the case?
3. Concept completion	Who? What? When? Where?
4. Feature specification	What qualitative properties does X have?
5. Quantification	How much? How many?
6. Definition questions	What does X mean?
7. Example questions	What is an example of a category?
8. Comparison	How is X similar to or different from Y?
9. Interpretation	What can be inferred from given data?
10. Causal antecedent	What state causally led to another state?
11. Causal consequence	What are the consequences of a state?
12. Goal orientation	What are the goals behind an agent action?
13. Procedural	What process allows an agent to reach a goal?
14. Enablement	What resource allows an agent to reach a goal?
15. Expectation	Why did some expected event not occur?
16. Judgmental	What value does the answerer give to an idea?
17. Assertion	A declarative statement that indicates the speaker does not understand an idea.
18. Request/Directive	The questioner wants the listener to perform some action.

et al. 1994). Intuitively, though questions are being asked by both student and tutor, the goal behind each question depends greatly upon who is asking it. For example, human tutor questions, unlike student questions, do not signify knowledge deficits and are instead attempts to facilitate student learning. Graesser et al. (1992) present an analysis of questions that occur during tutoring sessions that decouples the surface form of the question, the content of the question, the mechanism that generated the question, and the specificity of the question. Each of these are independent dimensions along which questions may vary. Although deep theoretical issues are addressed in the Graesser et al. (1992) analysis, in what follows we present the analysis as a descriptive or taxonomic framework to organize further discussion of human question generation.

2.1 Content

A question taxonomy can focus on surface features, such as the question stem used, or alternatively can focus on the conceptual content behind the question. As discussed by Graesser et al. (1992), there are several reasons to prefer a conceptual organization. One reason is that question stems underspecify the nature of the question. For example, "What happened" requires a causal description of events while "What is that" requires only a label or definition, even though both use the same stem, "What." Likewise questions can be explicitly marked with a question mark, but they can also be pragmatically implied, e.g. "I don't understand gravity." These distinctions motivate Table 1, which draws on previous research and has been validated in multiple studies (Graesser and Person 1994, Graesser et al. 1995, Person et al. 1994). In particular, question types 10 through 15 are highly correlated with the deeper levels of cognition in Bloom's taxonomy of educational objectives in the cognitive domain (Bloom 1956, Graesser and Person 1994). Thus one key finding along this dimension of analysis is that generation of optimal questions for student learning should be sensitive to conceptual content rather than merely surface form.

2.2 Mechanism

Graesser et al. (1992) specify four major question generation mechanisms. The first of these is knowledge deficit, which, being information seeking, generates learner questions more often than tutor questions (Graesser and Person 1994). The second mechanism is common ground. Questions

Table 2: Graesser, Person, and Huber (1992)'s Question Generation Mechanisms

Correction of Knowledge Deficit
1. Obstacle in plan or problem solving
2. Deciding among alternatives that are equally attractive
3. Gap in knowledge that is needed for comprehension
4. Glitch in explanation of an event
5. Contradiction
Monitoring Common Ground
6. Estimating or establishing common ground
7. Confirmation of a belief
8. Accumulating additional knowledge about a topic
9. Comprehension gauging
10. Questioner's assessment of answerer's knowledge
11. Questioner's attempt to have answerer generate an inference
Social Coordination of Action
12. Indirect request
13. Indirect advice
14. Asking permission
15. Offer
16. Bargaining
Control of Conversation and Attention
17. Greeting
18. Reply to summons
19. Change in speaker
20. Focus on agent's actions
21. Rhetorical question
22. Gripe

generated by this mechanism seek to maintain mutual knowledge and understanding between tutor and student, e.g. "Have you covered factorial designs?". The third mechanism coordinates social actions including requests, permission, and negotiation. Tutors ask these kinds of questions to engage the student in activities with pedagogical significance. The fourth and final question generation mechanism is conversation-control, by which the tutor and student affect the flow of the conversation and each other's attention, e.g. greetings, gripes, and rhetorical questions. Specific examples of the four question generation mechanisms are given in Table 2.

Arguably, the most important question generation mechanism for tutors is the common ground mechanism, accounting for more than 90% of tutor questions in two different domains, with the vast majority of these being student assessment questions (Graesser and Person 1994). Student assessment questions probe student understanding to confirm that it agrees with the tutor's understanding. Since the dimension of question content is independent of question mechanism, any of the questions in Table 1 can be generated as a student assessment question.

An interesting question for future research is the extent to which a question generated by a specific mechanism can have pragmatic effects consistent with other mechanisms, or alternatively, whether multiple mechanisms can be involved in the generation of a question. Intuitively, a tutor question can have multiple pragmatic effects on the student such as focusing attention, highlighting the importance of a topic, stimulating student self-assessment on that topic, assessing the student on that topic, creating an opportunity for student knowledge construction, or creating a context for further instruction (answer feedback, remediation, etc.). Some of these effects can be manifested by the surface realization of the question, which we turn to next.

2.3 Specificity

When the information sought by a question is explicitly marked, then that question is said to have a high degree of specificity. For example, "What is the first element of the list A, B, C?" is highly

Table 3: Question Specificity

Question Type	Specificity	Example
Pumps	Low	Can you say more?
Hints	Medium	What's going on with friction here?
Prompts	High	What's the force resisting the sliding motion of surfaces?

explicit and requires only knowledge of how to apply a *first* operator to a list. Contrastingly, "What's the first letter of the alphabet?" presupposes that the listener has the world knowledge and dialogue context to correctly identify the implied list, e.g. the Latin alphabet rather than the Cyrillic. Questions can be even less specified "What is the first letter?" or "What is it?" requiring greater degrees of common ground between the tutor and student.

Previous research has characterized specificity as being low, medium, or high to allow reliable coding for discriminative analyses (Graesser et al. 1992, Graesser and Person 1994, Person et al. 1994). However, it appears that low, medium, and high specificity questions map onto the kinds of questions that tutors ask in naturalistic settings known as pumps, hints, and prompts (Graesser et al. 1995, Hume et al. 1996) as is shown in Table 3. Questions at the beginning of the table provide less information to the student than questions towards the end of the table.

While questions generated via the student assessment mechanism described in Section 2.2 all have the consequence of creating an opportunity to assess student knowledge, there are several other possible effects. First, a tutorial strategy that asks more specific questions only when a student is floundering promotes active construction of knowledge (Graesser et al. 1995, Chi et al. 2001). Secondly, a very specific question, like a prompt, focuses attention on the word prompted for and highlights its importance (D'Mello et al. 2010). Thirdly, a less specific question can lead to an extended discussion, creating a context for further instruction (Chi et al. 2008). This list of effects is not meant to be exhaustive, but rather it highlights that there are many desirable effects that can be obtained by varying the specificity of tutorial questions.

3. Psychological Models

The psychological framework outlined in Section 2 has led to detailed psychological models of question asking. However, question asking in students has received more attention than in tutors because deep student questions are positively correlated with their test scores (Graesser et al. 1995). Thus there is considerable interest in scaffolding students to generate deep questions (Beck et al. 1997, Bransford et al. 1991, Brown 1988, Collins 1988, Dillon 1988, Edelson et al. 1999, Palinscar and Brown 1984, Piaget 1952, Pressley and Forrest-Pressley 1985, Scardamalia and Bereiter 1985, Schank 1999, Zimmerman 1989). Indeed there is an extensive literature investigating the improvements in the comprehension, learning, and memory of technical material that can be achieved by training students to ask questions during comprehension (Ciardiello 1998, Craig et al. 2006, Davey and McBride 1986, Foos 1994, Gavelek and Raphael 1985, King 1989, 1992, 1994, Odafe 1998, Palinscar and Brown 1984, Rosenshine et al. 1996, Singer and Donlan 1982, Wong 1985). Rosenshine et al. (1996) present a meta-analysis of 26 empirical studies investigating question generation learning effects.

A cognitive computational model of question asking has been developed by Graesser and colleagues (Graesser and Olde 2003, Otero and Graesser 2001). The model is called PREG, which is a root morpheme for "question" in the Spanish language. According to the PREG model, cognitive disequilibrium drives the asking of questions (Berlyne 1960, Chinn and Brewer 1993, Collins 1988, Festinger 1962, Flammer 1981, Graesser et al. 1996, Graesser and McMahen 1993, Schank 1999). Thus the PREG model primarily focuses on the knowledge deficit mechanisms of Table 2, and in its current form is most applicable to student-generated questions.

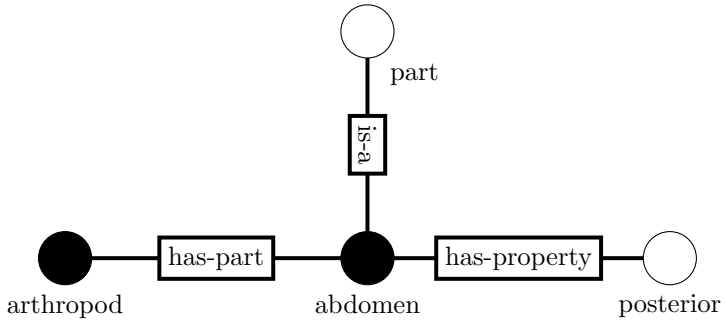

Figure 2: A concept map fragment. Key terms have black nodes.

PREG has two primary components. The first is a set of production rules that specify the categories of questions that are asked under particular conditions (i.e., content features of text and knowledge states of individuals). The second component is a conceptual graph, which is a particular instantiation of a semantic network (Graesser and Clark 1985, Sowa 1992). In this formulation of conceptual graphs, nodes themselves can be propositions, e.g. "a girl wants to play with a doll," and relations are (as much as possible) limited to a generic set of propositions for a given domain. For example, one such categorization consists of 21 relations including `is-a`, `has-property`, `has-consequence`, `reason`, `implies`, `outcome`, and `means` (Gordon et al. 1993). A particular advantage of limiting relations to these categories is that the categories can then be set into correspondence with the question types described in Table 1 for both the purposes of answering questions (Graesser and Franklin 1990) as well as generating them (Gordon et al. 1993).

In one study, the predictions of PREG were compared to the questions generated by middle and high school students reading science texts (Otero and Graesser 2001). The PREG model was able to not only account for nearly all of the questions the students asked, but it was also able to account for the questions the students didn't ask. The empirical evidence suggests that conceptual graph structures provide sufficient analytical detail to capture the systematic mechanisms of question asking.

4. Computational Model

The previous discussion in Sections 2 and 3 offers substantial guidance in the design of a computational model of tutorial question generation. Section 2 describes a conceptual question taxonomy in which deeper questions are correlated with deeper student learning, question generation mechanisms behind tutorial questions, and the role of specificity in question asking. Section 3 presents a psychological model, PREG, that builds on this framework. Although the PREG model is not completely aligned with the present objective of generating tutorial questions, it provides a roadmap for a two-step modeling approach. In the first step, conceptual graphs are created for a particular domain. Automatic extraction from text is desirable because hand-authoring knowledge representations for intelligent tutoring systems is extremely labor intensive (Murray 1998, Corbett 2002, Aleven et al. 2006). Secondly, conceptual graphs are used to generate student-assessment questions that span the question types of Table 1 and the specificity levels of Table 3.

We have previously presented an approach for extracting concept maps from a biology textbook (Olney 2010, Olney et al. 2010, 2011) that we will briefly review. Concept maps are similar to conceptual graph structures, but they are generally less structured (Fisher et al. 2000, Mintzes et al. 2005). Our concept map representation has two significant structural elements. The first is key terms, shown as black nodes in Figure 2. These are terms in our domain that are pedagogically

significant. Only key terms can be the start of a triple, e.g. *abdomen* `is-a` *part*. End nodes can contain key terms, other words, or complete propositions. The second central aspect of our representation is labeled edges, shown as boxes in Figure 2. As noted by Fisher et al. (2000), a small set of edges can account for a large percentage of relationships in a domain.

In comparison, the conceptual graph structures of Gordon et al. (1993) contain nodes that can be concepts or statements, and each node is categorized as a state, event, style, goal, or action. Their nodes may be connected with directional or bi-directional arcs from a fixed set of relations. However, conceptual graphs and our concept maps share a prescribed set of edges, and as will be discussed below, this feature facilitates linkages between the graph representations and question asking/answering (Graesser and Clark 1985, Graesser and Franklin 1990, Gordon et al. 1993). In the next few sections we provide an overview of the concept map extraction process.

4.1 Key Terms

As discussed by Vanderwende (2008), one goal of question generation systems should be to ask *important* questions. In order to do so reliably, one must identify the key ideas in the domain of interest. Note that at this level, importance is defined as generally important, rather than important with respect to a particular student. Defining important questions relative to a particular model of student understanding is outside the scope of this study.

Though general purpose key term extraction procedures have been proposed (Medelyan et al. 2009), they are less relevant in a pedagogical context where key terms are often already provided, whether in glossaries (Navigli and Velardi 2008), or textbook indices (Larrañaga et al. 2004). To develop our key terms, we used the glossary and index from a textbook in the domain of biology (Miller and Levine 2002) as well as the keywords given in a test-prep study guide (Cypress Curriculum Services 2008). This process yielded approximately 3,000 distinct terms, with 2,500 coming from the index and the remainder coming from the glossary and test-prep guide. Thus we can skip the keyword extraction step of previous work on concept map extraction (Valerio and Leake 2008, Zouaq and Nkambou 2009).

4.2 Edge Relations

Edge relations used in conceptual graphs typically depict abstract and domain-independent relationships (Graesser and Clark 1985, Gordon et al. 1993). However, previous work suggests that while a large percentage of relations in a domain are from a small set, new content can drive new additions to that set (Fisher et al. 2000). In order to verify the completeness of our edge relations, we undertook an analysis of concept maps from biology. We manually analyzed 4,371 biology triples available on the Internet[2]. These triples span the two topics of molecules & cells and population biology. Because these two topics represent the extremes of levels of description in biology, we presume that their relations will mostly generalize to the levels between them.

A frequency analysis of these triples revealed that 50% of all relations are `is-a`, `has-part`, or `has-property`. In the set of 4,371 biology triples, only 252 distinct relation types were present. We then manually clustered the 252 relation types into 20 relations. The reduction in relation types lost little information because the original data set had many subclassing relationships, e.g. `part-of` had the subclasses `composed of, has organelle, organelle of, component in, subcellular structure of`, and `has subcellular structure`. Since this subclassing is often recoverable by knowing the node type connected, e.g. an organelle, the mapping of the 252 relation types to 20 relations loses less information than might be expected. For example, weighted by frequency, 75% of relations in the 4,371 triples can be accounted for by the 20 relations by considering subclasses as above. If one additionally allows verbal relations to be subclassed by `has-consequence`, e.g. `eats` becomes `has-consequence` *eats X* where the subclassed verbal relation is merged with the end node

2. http://www.biologylessons.sdsu.edu

X, the 20 relations account for 91% of triples by frequency. Likewise applying `part-of` relationships to processes, as illustrated by *interphase* `part-of` *cell cycle* raises coverage to 95%. The remaining relations that don't fit well into the 20 clusters tend to be highly specific composites such as `often co-evolves with` or `gains phosphate to form`; these are subclassed under `has-property` in the same way as verbal relations are subclassed under `has-consequence`.

Of the 20 relations, 8 overlap with domain-independent relationships described in psychological research (Graesser and Clark 1985, Gordon et al. 1993) and an additional 4 are domain-general adjunct roles such as `not` and `direction`. The remaining 8 are biology and education specific, like `convert` and `definition`. These 20 relations were augmented with 10 additional relations from Gordon et al. and adjunct roles, including `after, contrast, enable, has-consequence, lack, produce, before, convert, example, has-part, location, purpose, combine, definition, extent, has-property, manner, reciprocal, connect, direction, follow, implies, not, require, contain, during, function, isa, possibility`, and `same-as`, in order to maximize coverage on unseen biology texts. A detailed analysis of the overlap between the clustered relations, relations from Gordon et al. (1993), and adjunct relations is presented in Olney et al. (2011).

4.3 Automatic Extraction

The automatic extraction process creates triples matching the representational scheme described in the previous sections. Since each triple begins with a key term, multiple triples starting with that term can be indexed according to that term. Alternatively, one can consider each key term to be the center of a radial graph of triples. Triples containing key terms as start and end nodes bridge these radial graphs.

Input texts for automatic extraction for the present study consisted of the textbook and glossary described in Section 4.1 as well as a set of 118 mini-lectures written by a biology teacher. Each of these mini-lectures is approximately 300-500 words long and is written in an informal style. The large number of anaphora in the mini-lectures necessitated using an anaphora resolution algorithm. Otherwise, triples that start with an anaphor, e.g. "It is contributing to the destruction of the rainforest," will be discarded because they do not begin with a key term.

Thus the first step in automatic extraction for the mini-lectures is anaphora resolution. EmPronoun is a state of the art pronoun anaphora resolution utility that has an accuracy of 68% on the Penn Treebank, roughly 10% better performance than previous methods implemented in JavaRAP, Open-NLP, BART and GuiTAR (Charniak and Elsner 2009). Using the EmPronoun algorithm, each lecture was rewritten by replacing anaphora with their respective reference, e.g. replacing "It" with "Logging" in the previous example.

The second step is semantic parsing. The LTH SRL[3] parser is a semantic role labeling parser that outputs a dependency parse annotated with PropBank and NomBank predicate/argument structures (Johansson and Nugues 2008, Meyers et al. 2004, Palmer et al. 2005). For each word token in a parse, the parser returns information about the word token's part of speech, lemma, head, and relation to the head. Moreover, it uses PropBank and NomBank to identify predicates in the parse, either verbal predicates (PropBank) or nominal predicates (NomBank), and their associated arguments. For example, consider the sentence, "Many athletes now use a dietary supplement called creatine to enhance their performance." The LTH SRL parser outputs five predicates for this sentence:

use (athletes/A0) (now/AM-TMP) (supplements/A1) (to/A2)

supplement (dietary/A1) (supplement/A2)

called (supplement/A1) (creatine/A2)

enhance (supplement/A0) (performance/A1)

3. The Swedish "Lunds Tekniska Högskola" translates as "Faculty of Engineering."

Table 4: Example Predicate Maps

Predicate	POS	Edge Relation	Frequency	Start	End
have.03	V	HAS_PROPERTY	1,210	A0	Span
use.01	V	USE	1,101	A0	Span
produce.01	V	PRODUCE	825	A0	Span
call.01	V	HAS_DEFINITION	663	A1	A2

performance (their/A0)

Three of these predicates are verbal predicates: "use," "called," and "enhance." Verbal predicates often, but not always, have agent roles specified by A0 and patient roles specified by A1. However, consistent generalizations in role labeling are often lacking, particularly for arguments beyond A0 and A1 (Palmer et al. 2005). Moreover, the presence of A0 and A1 can't be guaranteed: although "use" and "enhance" both have A0 and A1, "called" has no A0 because there's no clear agent doing the calling in this situation. Passive verbs frequently have no A0. Finally, verbal adjuncts can specify adverbial properties such as time, here specified by AM-TMP.

Two of the predicates are nominal predicates: "supplement" and "performance." Nominal predicates have much more complex argument role assignment rules than verbal predicates (Meyers et al. 2004). Predicates that are nominalizations of verbs, such as "performance," have roles more closely corresponding to verbal predicate roles like A0 and A1. For other kinds of nominal predicates, overt agents taking A0 roles are less common. As a result, many nominal predicates have only an A1 filling a theme role, as "dietary" is the theme of supplement.

The third step in concept map extraction is the actual extraction step. We have defined four extractor algorithms that target specific syntactic/semantic features of the parse, `is-a`, adjectives, prepositions, and predicates. Each extractor begins with an attempt to identify a key term as a possible start node. Since key terms can be phrases, e.g. "homologous structure," the search for key terms greedily follows the syntactic dependents of a potential key term while applying morphological rules. For example, in the sentence "Homologous structures have the same structure," "structures" is the subject of the verb "have" and a target for a key term. The search process follows syntactic dependents of the subject to map "homologous structures" to the known key term "homologous structure." In many cases, no key term will be found, so the prospective triple is discarded.

Several edge relations are handled purely syntactically. `Is-a` relations are indicated when the root verb of the sentence is "be," but not a helping verb. `Is-a` relations can create a special context for processing additional relations. For example, in the sentence, "An abdomen is a posterior part of an arthropod's body," "posterior" modifies "part," but the desired triple is *abdomen* `has-property` *posterior*. This is an example of the adjective extraction algorithm running in the context of an `is-a` relation. Prepositions can create a variety of edge relations. For example, if the preposition is IN and has a LOC dependency relation to its head (a locative relation), then the appropriate relation is *location*, e.g. "by migrating whales in the Pacific Ocean" becomes *whales* `location` *in the Pacific Ocean*.

Relations from PropBank and NomBank require a slightly more sophisticated approach. As illustrated in some of the preceding examples, not all predicates have an A0. Likewise not all predicates have patient/instrument roles like A1 and A2. The variability in predicate arguments makes simple mapping, e.g. A0 is the start node, predicate is the edge relation, and A1 is the end node, unrealistic. Therefore we created a manual mapping between predicates, arguments, and edge relations for every predicate that occurred more than 40 times in the corpus (358 predicates total). Table 4 lists the four most common predicates and their mappings.

The label "Span" in the last column indicates that the end node of the triple should be the text dominated by the predicate. For example, "Carbohydrates give cells structure." has AO "carbohydrates" and A1 "structure." However, it is more desirable to extract the triple *carbohydrates*

`has-property` *give cells structure* than *carbohydrates* `has-property` *structure*. End nodes based on predicate spans tend to contain more words and therefore have closer fidelity to the original sentence.

Finally, we apply some filters to remove triples that are either not particularly useful for question generation or appear to be from mis-parsed sentences. We apply filters on the back end rather than during concept map extraction because some of the filtered triples are useful for other purposes besides question generation, e.g. student modeling and assessment. Three of the filters are straightforward and require little explanation: the repetition filter, the adjective filter, and the nominal filter. The repetition filter considers the number of words in common between the start and end nodes. If the number of shared words is more than half the words in the end node, the triple is filtered. This helps alleviate redundant triples such as *cell* `has-property` *cell*. The adjective filter removes any triple whose key term is an adjective. These triples violate the assumption by the question generator that all key terms are nouns. For example, 'reproducing' might be a key term, but in a particular triple it might function as an adjective rather than a noun. These cases are often caused by mis-parsed sentences. Finally, the nominal filter removes all NomBank predicates except *has-part* predicates, because these often have Span end nodes and so contain themselves, e.g. *light* `has-property` *the energy of sunlight*.

The most sophisticated filter is the likelihood ratio filter. This filter measures the association between the start and end node using likelihood ratios (Dunning 1993) and a chi-square significance criterion to remove triples with insignificant association. Words from the end node that have low log entropy are removed prior to calculation, and the remaining words from start and end nodes are pooled. The significance criterion for the chi-square test is .0001. This filter weeds out start and end nodes that do not have a strong association.

In the document set used for the present study, 43,833 of the originally extracted triples were filtered to a set of 19,143 triples. The filtered triples were distributed around 1,165 start nodes out of approximately 3,000 possible key terms. The five most connected key terms in the filtered set are cell, plant, organism, species, and animal, which collectively account for 18% of the total connections. Of the possible 30 edge relations, only 18 were present in the filtered triples, excluding `before, convert, direction, during, extent, follow, function, implies, manner, possibility, reciprocal,` and `after`. The top five edge relations extracted were `has-property, has-consequence, has-part, location`, and `is-a`, making up roughly 82% of the total relations. By themselves, the relations `has-property, is-a`, and `has-part` make up 56% of all edge relations, which is consistent with human concept maps for biology domains (Fisher et al. 2000).

4.4 Question Generation

Our question generation approach uses the concept map described in previous sections to generate questions. Questions may either be generated from individual triples or by combinations of triples. Question generation from individual triples is very similar to generating questions from individual sentences. Both cases ignore how information is structured across the domain. Question generation from combinations of triples, in contrast, introduces a limited form of reasoning over the concept map knowledge representation. We consider each of these in turn.

4.4.1 Generation from Individual Triples

Question generation using individual triples can generate questions of varying specificity as described in Section 2.3. In general, the less specific a question is, the easier it is to generate the question with a template. For example, pumps, the least specific question type, can be generated with a fixed bag of expressions like, "What else?" or "Can you say more?" Although this level of specificity is trivial, the generation of hints and prompts warrants some discussion.

Hints have an intermediate level of specificity. Our approach to hint generation makes use of parametrized question templates that consider the start node and edge relation of a triple. Some example hint templates of the 14 that were used are given in Table 5. Question templates are

Table 5: Example Hint Question Templates

Edge Relation	Question Template
?	And what do we know about **KT**?
HAS_CONSEQUENCE	What *do* **KT** do?
HAS_DEFINITION	**KT**, what is that?
HAS_PART	What *do* **KT** have?
ISA	So **KT** *be*?

selected based on the edge relation of the source triple, or a wildcard template (?) can be used. Templates may have placeholders for the key term (**KT**) as well as verb lemmas (*do, be*). Question templates for hints are populated in two steps. First, a determiner is added to the key term. The algorithm decides what determiner to add based on whether a determiner modified the key term in the source sentence, the key term is a mass noun, or the key term is plural. The second step involves matching the agreement features of the key term with verb lemma (if present). Agreement features are derived using the SPECIALIST Lexicon (Browne et al. 2000), which has been previously used in the natural language generation community (Gatt and Reiter 2009).

Prompts have the highest level of specificity, because by definition prompts seek only a word or phrase for an answer (Graesser et al. 1995). Additionally, prompts are often generated as incomplete assertions (see Table 1), e.g. "The first letter of the alphabet is...?" By making use of this assertion-oriented structure as well as the single relations encoded in triples, our prompt generator attempts to avoid extremely complex questions that can be created by syntactic transformation, e.g. "Where was penicillin discovered by Alexander Fleming in 1928?"

Prompt generation is a two step process. In the first step, the triple is rendered as an assertion. Rendering an assertion requires assembling the start node, edge relation, and end node into a coherent sentence. Constructing a well formed declarative sentence from these fragments requires managing a large number of possible cases involving tense, determiners, modifiers, passivization, and conjunction, amongst others. The assertion generation process can be considered as a black box for the purposes of the current discussion.

If the edge relation is `not`, then the declarative sentence is transformed into a verification question. Otherwise the declarative sentence is scanned for the word with the highest log entropy weight (Dumais 1991). This word is substituted by "what" to make the final prompt. By using log entropy as a criterion for querying, we are attempting to maximize the probability that the deleted item is relevant in the domain. Log entropy also gives a principled way of selecting amongst multiple key terms in a sentence, though in our implementation any word was a target.

4.4.2 Generation from Multiple Triples

One significant advantage to building a knowledge representation of text prior to generating questions is that knowledge may be integrated across the text. In our model, integration naturally falls out of the restriction that all triples must start with a key term. As long as triples begin with the same key term, they may be considered as relating to the same concept even if they are many pages apart. The additional structure this provides allows for some interesting questioning strategies. We briefly define three questioning strategies that we call contextual verification, forced choice, and causal chain, respectively.

Both contextual verification questions and forced choice questions compare and contrast features of two related concepts. Empirical evidence suggests that this kind of discriminative learning, in which salient features are compared and contrasted across categories, is an important part of concept learning (Tennyson and Park 1980). There are some similarities between this strategy and the Teaching with Analogies philosophy (Glynn 2008). In both cases, the student is reminded of some-

thing they know in order to learn something new. The context primes existing student knowledge and helps the student associate that knowledge with the situation presented in the question.

Contextual verification questions present a context and then ask a verification question. An example contextual verification question from our system is "An ecosystem is a community. Is that true for a forest?" In order to generate contextual verification questions, we index all of the key nodes that are subtypes (`is-a`) of a common node. For example, *cat* and *dog* are both subtypes of *animal*. Each key node has an associated set of triples that can be intersected with the other sets to discover common and unique properties. For example, cats and dogs both have tails and four legs, but only cats chase mice. Common properties can be used to generate contextual verification questions that should be answered positively, e.g. "Cats have a tail. Is that true for dogs?" while unique properties can be used to generate questions that should be answered negatively "Cats chase mice. Is that true for dogs?" The context component is generated as an assertion as described previously for prompts. The verification question itself is easily generated using a question template as described previously for hints.

Forced choice questions in contrast can only apply to the disjunctive case in which a property is not held in common by two subtypes of a common node. An example generated by our system is "What resides in skin, melanin or phytochrome?" The triples used to generate forced choice questions are selected using the same method as for contextual verification questions, and likewise may be generated using a question template.

Causal chain questions are based on causal relationships that connect multiple key terms. Previous research has investigated how students can learn an underlying causal concept map and use it to solve problems and construct explanations (Mills et al. 2004). Qualitatively speaking, causal relations have a +/- valence, indicating a direct or inverse relationship of one concept/variable to another. For example, carrots may be directly causally related (+) to rabbits, such that an increase in the number of carrots leads to an increase in the number of rabbits. Causal chain questions are constructed by joining two triples from the concept map, such that the end of the first triple is the same as the beginning of the second triple. An example causal chain question produced by our system is "How do bacteria produce energy?" which bridges the key terms of *bacteria*, *food*, and *energy*. Since causal chaining requires the edge relations be transitive, we restricted causal chaining to the edge relations `produce` and `has-consequence`.

5. Evaluation

5.1 Method

We conducted an evaluation of the question generation system described in Section 4. Three judges who were experts in questions and pedagogy evaluated questions generated by the system. Each question was rated on the following five dimensions: whether the sentence was of the target type (*Question Type*), the relevance of the question to the source sentence (*Relevance*), the syntactic fluency of the question (*Fluency*), the ambiguity of the question (*Ambiguity*), and the pedagogical value of the question (*Pedagogy*). The first four of these dimensions were derived from the Question Generation Shared Task Evaluation Challenge (QGSTEC) (Rus et al. 2010b; Rus et al., this volume). A consistent four item scale was used for all dimensions except question type, which was binary. An example of the four item scale is shown in Table 6. The full set of scales is included in the Appendix.

The evaluation set rated by the judges was constructed using the three text sources described in Section 4.3: the textbook, glossary, and mini-lectures. Each judge blindly rated the same 30 hint and prompt questions from each of these sources, for a total of 60 questions from each source. In addition, each judge rated approximately 30 contextual verification, forced choice, and causal chain questions. Since these generation methods are hard hit by sparse data, all three text sources were used to generate these three question types. Each hint or prompt was preceded by its source sentence so that comparative judgments, like relevance, could be made:

Table 6: Rating Scale for Relevance

Score	Criteria
1	The question is completely relevant to the input sentence.
2	The question relates mostly to the input sentence.
3	The question is only slightly related to the input sentence.
4	The question is totally unrelated to the input sentence.

Table 7: Inter-rater Reliability

Scale	Judge Pair	Cronbach's α
Question Type	AB	.43
Relevance	AB	.82
Fluency	AB	.79
Ambiguity	AB	.74
Pedagogy	BC	.80

An antheridium is a male reproductive structure in some algae and plants.
Tell me about an antheridium. (**Hint**)
An antheridium is what? (**Prompt**)

The contextual verification, forced choice, and causal chain questions were preceded by the two source sentences for their respective triples:

Enzymes are proteins that act as biological catalysts.
Hemoglobin is the protein in red blood cells that carries oxygen.
Enzymes are proteins. Is that true for hemoglobin?

Inter-rater reliability was calculated on each of the five measures, using a two-way random effect model to calculate average measure intra-class correlation. Cronbach's α for each measure is presented in Table 7. The pair of judges with the highest alpha for each measure were used to calculate composite scores for each question in later analyses. Most of the reliability scores in Table 7 are close to .80, which is considered satisfactory reliability. However, reliability for question type was poor at $\alpha = .43$. This smaller value is attributable to the fact that these judgments were binary and to the conservative test statistic: proportion agreement for question type was .80 for hints and .75 for prompts.

5.2 Results

Our two guiding hypotheses in the evaluation were that question source and question type would affect ratings scores. Question source is likely to affect ratings because sentences from the texts vary in terms of their syntactic complexity. For example, sentences from the glossary have the prototype structure "An X is a Y that ..." while sentences from the other sources have no such restrictions. Additionally, anaphora resolution was used on mini-lectures but not on the other texts. This could affect questions generated from mini-lectures negatively by introducing errors in anaphora resolution or positively by removing ambiguous anaphora from questions. Likewise question categories vary considerably in the complexity of processes used to generate them. Hints are more template based than prompts, and both of these question types are simpler than contextual verification, forced choice, and causal chain questions that construct questions over multiple sentences. In the following sections we present statistical tests of significance between these conditions.

Table 8: Mean Ratings Across Question Sources

	Textbook		Mini-lectures		Glossary	
Scale	Mean	SD	Mean	SD	Mean	SD
Question Type	1.41	0.36	1.44	0.38	1.27	0.42
Relevance	2.72	1.02	2.54	0.92	1.77	1.08
Fluency	2.29	1.01	2.17	0.98	2.25	1.14
Ambiguity	3.35	0.89	3.19	0.88	2.79	0.82
Pedagogy	2.88	1.07	3.09	1.03	2.25	1.18

5.2.1 Results of Question Source

A one-way ANOVA was used to test for relevance differences among the three question sources. Relevance differed significantly across the three sources, $F(2,177) = 14.80$, $p = .0001$. The effect size, calculated using Cohen's f^2, was .17. Scheffé post-hoc comparisons of the three sources indicate that the glossary questions ($M = 1.78$, $SD = 1.08$) had significantly better relevance than the textbook questions ($M = 2.72$, $SD = 1.02$), $p = .0001$ and significantly better relevance than the mini-lecture questions ($M = 2.54$, $SD = 0.92$), $p = .0001$.

This pattern was repeated for ambiguity, $F(2,177) = 6.65$, $p = .002$. The effect size, calculated using Cohen's f^2, was .08. Scheffé post-hoc comparisons indicate that the glossary questions ($M = 2.79$, $SD = 0.82$) had significantly lower ambiguity than the textbook questions ($M = 3.35$, $SD = 0.89$), $p = .002$ and the mini-lecture questions ($M = 3.19$, $SD = 0.88$), $p = .043$.

The same pattern held for pedagogy, $F(2,177) = 9.60$, $p = .0001$. The effect size, calculated using Cohen's f^2, was .11. Scheffé post-hoc comparisons indicate that the glossary questions ($M = 2.25$, $SD = 1.18$) had significantly better pedagogy than the textbook questions ($M = 2.88$, $SD = 1.08$), $p = .008$ and the mini-lecture questions ($M = 3.09$, $SD = 1.03$), $p = .0001$.

One-way ANOVAs were used to test for question type and fluency differences among the three question sources, but neither question type nor fluency significantly differed across the three sources of questions, $p = .05$. Mean ratings across question sources are presented in Table 8.

5.2.2 Results of Question Category

A one-way ANOVA was used to test for question type differences across the five question categories. Question type differed significantly across the five categories, $F(4,263) = 12.00$, $p = .0001$. The effect size, calculated using Cohen's f^2, was .18. Scheffé post-hoc comparisons of the five question categories indicate that the prompt questions ($M = 1.55$, $SD = 0.42$) were significantly less likely to be of the appropriate type than the hint questions ($M = 1.2$, $SD = 0.26$), $p = .0001$ and significantly less likely to be of the appropriate type than the forced choice questions ($M = 1.27$, $SD = 0.37$), $p = .009$.

A one-way ANOVA was used to test for fluency differences across the five categories. Fluency differed significantly across the five categories, $F(4,263) = 29.40$, $p = .0001$. The effect size, calculated using Cohen's f^2, was .45. Scheffé post-hoc comparisons of the five question categories indicate that the hint questions ($M = 1.56$, $SD = 0.74$) were significantly more fluent than prompts ($M = 2.92$, $SD = 0.83$), $p = .0001$, forced choice questions ($M = 2.64$, $SD = 1.04$), $p = .0001$, contextual verification questions ($M = 2.25$, $SD = 1.02$), $p = .007$, and causal chain questions ($M = 2.40$, $SD = 0.98$), $p = .0001$. Post-hoc comparisons further indicated that contextual verification questions were significantly more fluent that prompts, $p = .01$.

A one-way ANOVA was used to test for pedagogy differences across the five question categories. Pedagogy differed significantly across the five categories, $F(4,263) = 8.10$, $p = .0001$. The effect size, calculated using Cohen's f^2, was .12. Scheffé post-hoc comparisons of the five question categories indicate that the hint questions ($M = 2.39$, $SD = 1.15$) were significantly more pedagogic than prompts ($M = 3.09$, $SD = 1.03$), $p = .001$, forced choice questions ($M = 3.21$, $SD = 1.07$), $p = .015$,

Table 9: Mean Ratings for Single Triple Question Types

	Prompt		Hint	
Scale	Mean	SD	Mean	SD
Question Type	1.55	0.42	1.20	0.26
Relevance	2.45	1.12	2.24	1.04
Fluency	2.92	0.83	1.56	0.74
Ambiguity	3.12	0.97	3.10	0.81
Pedagogy	3.09	1.03	2.39	1.15

Table 10: Mean Ratings for Multiple Triple Question Types

	Forced Choice		Contextual Verification		Causal Chain	
Scale	Mean	SD	Mean	SD	Mean	SD
Question Type	1.27	0.37	1.42	0.37	1.37	0.29
Relevance	2.52	0.88	2.88	0.95	2.13	0.63
Fluency	2.64	1.04	2.25	1.02	2.40	0.98
Ambiguity	3.07	0.89	3.13	0.83	2.85	0.77
Pedagogy	3.21	1.07	3.33	1.02	3.18	1.05

contextual verification questions (M = 3.33, SD = 1.02), p = .002, and causal chain questions (M = 3.18, SD = 1.05), p = .018.

One-way ANOVAs were used to test for relevance and ambiguity differences among the five question categories, but neither relevance nor ambiguity significantly differed across the five question categories, p = .05. Mean ratings across question categories are presented in Table 9 and Table 10.

6. Discussion

Ideally, the results in Section 5.2 would be interpreted relative to previous results in the literature. However, since no official results from the 2010 Question Generation Shared Task Evaluation Challenge (QGSTEC) appear to have been released, comparison to previous results is rather limited. The only work that reports informal results using the QGSTEC scales appears to be the WLV system (Varga and Ha 2010). In this study, inter-annotator agreement for relevance and fluency in Cohen's κ is .21 and .22, for average ratings of 2.65 and 2.98 respectively. In comparison, our agreement on these scales in κ was .82 and .86. As no other inter-annotator agreement appears on the QGSTEC scales appears to have been reported elsewhere, one contribution of our study is to show that high reliability can be achieved on these five scales.

Although our range of relevance, 2.13-2.88, and our range of fluency, 1.56-2.92, appear to compare favorably to those of Varga and Ha (2010), it's not clear which of our question categories (if any) map onto theirs. Therefore it is more informative to discuss the comparisons of question source and question category and reflect on how these comparisons might inform our psychologically-based model of question generation.

Overall, most means in Tables 8, 9, and 10 fall in the indeterminate range between slightly correct and slightly incorrect. Notable exceptions to this trend are pedagogy for multiple triple categories and prompts, which tend towards slightly unreasonable, and ambiguity scores across tables that tend to slightly ambiguous. Encouragingly, these few means are closer in score indicating they are slightly unreasonable rather than completely unreasonable. One sample t-tests confirm that all these means are statistically different from completely unreasonable, p = .05. This is particularly encouraging for the question categories that make use of multiple triples, contextual verification, forced choice, and causal chain, because they have more opportunities to make errors.

The comparisons between question sources in Section 5.2.1 support the conclusion that our model converts glossary sentences into better questions than sentences from the textbook or mini-lectures, although the effect sizes are small, with f^2 ranging from .08-.17. This result supports the hypothesis that glossary sentences, having less syntactic variability than sentences from the other texts, don't require as complex algorithms to generate questions. Glossary sentences produce acceptably relevant questions, with a mean score of 1.77. On the other dimensions, glossary sentences are roughly in the middle of the scale between somewhat correct and somewhat incorrect.

It is somewhat surprising that there were no differences between the textbook and the mini-lectures, because the mini-lectures used the EmPronoun anaphora resolution algorithm (Charniak and Elsner 2009). Anaphora resolution errors have been previously reported as significant detractors in question generation systems that use knowledge representations (Chen 2009, Chen et al. 2009, Mostow and Chen 2009). However, these previous studies did not use EmPronoun, which has roughly 10% better performance than other state of the art methods (Charniak and Elsner 2009). Since no significant differences were found in our evaluation, we tentatively conclude that the EmPronoun algorithm, by resolving out-of-context pronouns, boosted scores as much as it hurt them.

The comparison between question categories in Section 5.2.2 is perhaps the richest and most interesting comparison in our evaluation. The first main finding of the question category comparison is that prompt questions are less likely to be of the correct type than hints or forced choice. The similarity in ambiguity ratings between prompts (M = 3.12) and hints (M = 3.10) suggests that prompts are not very specific and might be more like hints than prompts. The second major finding in Section 5.2.2 is that hints are more fluent and have better pedagogy than all the other categories. The success of hints in this regard is probably due to their template-based generation, which is the simplest question generation method of the five methods evaluated.

However, it's also worth noting what we didn't find, which are additional differences between the questions generated from multiple triples and those generated from a single triple. All three multiple triple question categories in Table 10 have neutral fluency and relevance scores, suggesting that they are on average acceptable questions. Additionally, no differences were found between single triple and multiple triple question categories for the measures of relevance and ambiguity. On the other hand, pedagogy scores for multiple triple question categories are less reasonable, indicating that more work needs to be done before these questions can be used in an e-learning environment.

7. Conclusion

The major goal of this study was to bridge the gap between psychological theories of question asking and computational models of question generation. The model we presented in Section 4 is heavily grounded in several decades of psychological research on question asking and answering. Our objective is not to build linkages to psychological theory merely for the sake of it. On the contrary, we believe that the psychological mechanisms that have been proposed for question asking hold great promise for future computational models of question generation. By generating plausible questions from concept maps, our model lends some support to this claim.

The computational model we presented in this study is one step towards the larger goal of understanding what question to generate, dynamically, for an individual student. However, many additional questions need to be answered before this goal can be realized. We must track what the student knows and use pedagogical knowledge to trigger the right questions at the right time. Solving these research challenges will deepen our understanding and advance the art and science of question generation.

Appendix A. Modified QGSTEC Rating Scales

Relevance

1 The question is completely relevant to the input sentence.

2 The question relates mostly to the input sentence.

3 The question is only slightly related to the input sentence.

4 The question is totally unrelated to the input sentence.

Question Type

1 The question is of the target question type.

2 The type of the generated question and the target question type are different.

Fluency

1 The question is grammatically correct and idiomatic/natural.

2 The question is grammatically correct but does not read as fluently as we would like.

3 There are some grammatical errors in the question.

4 The question is grammatically unacceptable.

Ambiguity

1 The question is completely unambiguous.

2 The question is mostly unambiguous.

3 The question is slightly ambiguous.

4 The question is totally ambiguous when asked out of the blue.

Pedagogy

1 Very reasonable

2 Somewhat reasonable

3 Somewhat unreasonable

4 Very unreasonable

References

Vincent Aleven, Bruce M. McLaren, Jonathan Sewall, and Kenneth R. Koedinger. The cognitive tutor authoring tools (CTAT): Preliminary evaluation of efficiency gains. In *Intelligent Tutoring Systems*, pages 61–70, 2006.

Husam Ali, Yllias Chali, and Sadid A. Hasan. Automation of question generation from sentences. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 58–67, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Isabel L. Beck, Margaret G. McKeown, Rebecca L. Hamilton, and Linda Kucan. *Questioning the Author: An Approach for Enhancing Student Engagement with Text.* International Reading Association, 1997.

Daniel Ellis Berlyne. *Conflict, Arousal, and Curiosity.* McGraw-Hill New York, 1960.

Benjamin S. Bloom, editor. *Taxonomy of educational objectives. Handbook I: Cognitive domain.* McKay, New York, 1956.

John D. Bransford, Susan R. Goldman, and Nancy J. Vye. Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. J. Sternberg and L. Okagaki, editors, *Influences on children*, pages 147–180. Erlbaum, Hillsdale, NJ, 1991.

Ann L. Brown. Motivation to learn and understand: On taking charge of one's own learning. *Cognition and Instruction*, 5:311–321, 1988.

Allen C. Browne, Alexa T. McCray, and Suresh Srinivasan. The Specialist Lexicon. Technical report, National Library of Medicine, Bethesda, Maryland, June 2000.

Xavier Carreras and Lluís Màrquez. Introduction to the conll-2004 shared task: Semantic role labeling. In Hwee Tou Ng and Ellen Riloff, editors, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.

Eugene Charniak and Micha Elsner. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece, March 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E09-1018`.

Wei Chen. Understanding mental states in natural language. In *Proceedings of the Eighth International Conference on Computational Semantics*, IWCS-8 '09, pages 61–72, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-90-74029-34-6. URL `http://portal.acm.org/citation.cfm?id=1693756.1693766`.

Wei Chen, Gregory Aist, and Jack Mostow. Generating questions automatically from informational text. In Scotty D. Craig and Darina Dicheva, editors, *The 2nd Workshop on Question Generation*, volume 1 of *AIED 2009 Workshops Proceedings*, pages 17–24, Brighton, UK, July 2009. International Artificial Intelligence in Education Society.

Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. Learning from tutoring. *Cognitive Science*, 25:471–533, 2001.

Michelene T. H. Chi, Marguerite Roy, and Robert G. M. Hausmann. Observing tutorial dialogues collaboratively: Insights about human tutoring effectiveness from vicarious learning. *Cognitive Science*, 32(2):301–341, 2008.

Clark A. Chinn and William F. Brewer. The Role of Anomalous Data in Knowledge Acquisition: A Theoretical Framework and Implications for Science Instruction. *Review of Educational Research*, 63(1):1–49, 1993.

Angelo V. Ciardiello. Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy*, 42(3):210–219, 1998.

Alan Collins. Different goals of inquiry teaching. *Questioning Exchange*, 2:39–45, 1988.

Alan Collins and Dedre Gentner. A framework for a cognitive theory of writing. In L. Gregg and E. R. Steinberg, editors, *Cognitive Processes in Writing*, pages 51–72. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1980.

Albert Corbett. Cognitive tutor algebra I: Adaptive student modeling in widespread classroom use. In *Technology and Assessment: Thinking Ahead. Proceedings from a Workshop*, pages 50–62, Washington, D.C., November 2002. National Academy Press.

Scotty D. Craig, Jeremiah Sullins, Amy Witherspoon, and Barry Gholson. The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction*, 24(4):565–591, 2006.

LLC Cypress Curriculum Services. *Tennessee Gateway Coach, Biology.* Triumph Learning, New York, NY, 2008.

Robert Dale, Barbara Di Eugenio, and Donia Scott. Introduction to the special issue on natural language generation. *Computational Linguistics*, 24(3):345–354, September 1998.

Beth Davey and Susan McBride. Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, 78(4):256–262, 1986.

James Thomas Dillon. *Questioning and Teaching. A Manual of Practice.* Teachers College Press, New York, 1988.

Sidney D'Mello, Patrick Hays, Claire Williams, Whitney Cade, Jennifer Brown, and Andrew M. Olney. Collaborative lecturing by human and computer tutors. In *Intelligent Tutoring Systems*, Lecture Notes in Computer Science, pages 178–187, Berlin, 2010. Springer.

Susan Dumais. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236, 1991.

Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, March 1993. ISSN 0891-2017. URL `http://portal.acm.org/citation.cfm?id=972450.972454`.

Daniel C. Edelson, Douglas N. Gordin, and Roy D. Pea. Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences*, 8(3 & 4): 391–450, 1999.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3), 2010. URL `http://www.aaai.org.proxy.lib.sfu.ca/ojs/index.php/aimagazine/article/view/2303`.

Leon Festinger. *A Theory of Cognitive Dissonance.* Tavistock Publications, 1962.

Kathleen M. Fisher, James H. Wandersee, and David E. Moody. *Mapping biology knowledge.* Kluwer Academic Pub, 2000.

August Flammer. Towards a theory of question asking. *Psychological Research*, 43(4):407–420, 1981.

Paul W. Foos. Student study techniques and the generation effect. *Journal of Educational Psychology*, 86(4):567–76, 1994.

Albert Gatt and Ehud Reiter. Simplenlg: a realisation engine for practical applications. In *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, pages 90–93, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

James R. Gavelek and Taffy E. Raphael. Metacognition, instruction, and the role of questioning activities. In D.L. Forrest-Pressley, G.E. MacKinnon, and G.T. Waller, editors, *Metacognition, cognition, and human performance: Instructional practices*, volume 2, pages 103–136. Academic Press, Orlando, FL, 1985.

Shawn M. Glynn. Making science concepts meaningful to students: teaching with analogies. In S. Mikelskis-Seifert, U. Ringelband, and M. Bruckmann, editors, *Four decades of research in science education: From curriculum development to quality improvement*, pages 113–125. Waxmann, Mnster, Germany, 2008.

Sallie E. Gordon, Kimberly A. Schmierer, and Richard T. Gill. Conceptual graph analysis: Knowledge acquisition for instructional system design. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35(3):459–481, 1993.

Arthur C. Graesser and Leslie C. Clark. *Structures and procedures of implicit knowledge*. Ablex, Norwood, NJ, 1985.

Arthur C. Graesser and Stanley P. Franklin. Quest: A cognitive model of question answering. *Discourse Processes*, 13:279–303, 1990.

Arthur C. Graesser and Cathy L. McMahen. Anomalous information triggers questions when adults solve problems and comprehend stories. *Journal of Educational Psychology*, 85:136–151, 1993.

Arthur C. Graesser and Brent A. Olde. How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, 95(3):524–536, 2003.

Arthur C. Graesser and Natalie K. Person. Question asking during tutoring. *American Educational Research Journal*, 31(1):104–137, 1994. ISSN 0002-8312.

Arthur C. Graesser, Sallie E. Gordon, and Lawrence E. Brainerd. Quest: A model of question answering. *Computers and Mathematics with Applications*, 23:733–745, 1992.

Arthur C. Graesser, Natalie K. Person, and Joseph P. Magliano. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology*, 9:1–28, 1995.

Arthur C. Graesser, William Baggett, and Kent Williams. Question-driven explanatory reasoning. *Applied Cognitive Psychology*, 10:S17–S32, 1996.

Michael Heilman and Noah A. Smith. Ranking automatically generated questions as a shared task. In Scotty D. Craig and Darina Dicheva, editors, *The 2nd Workshop on Question Generation*, volume 1 of *AIED 2009 Workshops Proceedings*, pages 30–37, Brighton, UK, July 2009. International Artificial Intelligence in Education Society.

Michael Heilman and Noah A. Smith. Extracting simplified statements for factual question generation. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 11–20, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Denis J. Hilton. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1): 65–81, 1990.

Gregory Hume, Joel Michael, Allen Rovick, and Martha Evens. Hinting as a tactic in one-on-one tutoring. *The Journal of the Learning Sciences*, 5:23–47, 1996.

Richard Johansson and Pierre Nugues. Dependency-based syntactic-semantic analysis with PropBank and NomBank. In *CoNLL '08: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 183–187, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-48-4.

Saidalavi Kalady, Ajeesh Elikkottil, and Rajarshi Das. Natural language question generation using syntax and keywords. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 1–10, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Alison King. Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology*, 14(4):1–16, 1989.

Alison King. Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, 29(2):303–323, 1992.

Alison King. Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31(2):338–368, 1994.

Mikel Larrañaga, Urko Rueda, Jon A. Elorriaga, and Ana Arruarte Lasa. Acquisition of the domain structure from document indexes using heuristic reasoning. In *Intelligent Tutoring Systems*, pages 175–186, 2004.

Prashanth Mannem, Rashmi Prasad, and Aravind Joshi. Question generation from paragraphs at UPenn: QGSTEC system description. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 84–91, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Danielle S. McNamara and Joseph P. Magliano. Toward a comprehensive model of comprehension. In Brian H. Ross, editor, *The psychology of learning and motivation*, volume 51, chapter 9, pages 297–384. Academic Press, New York, 2009. ISBN 978-0-12-374489-0.

Olena Medelyan, Eibe Frank, and Ian H. Witten. Human-competitive tagging using automatic keyphrase extraction. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1318–1327, Singapore, August 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/D/D09/D09-1137`.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. The NomBank project: An interim report. In A. Meyers, editor, *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Kenneth R. Miller and Joseph S. Levine. *Prentice Hall Biology.* Pearson Education, New Jersey, 2002.

Bruce Mills, Martha Evens, and Reva Freedman. Implementing directed lines of reasoning in an intelligent tutoring system using the Atlas Planning Environment. In *Proceedings of the International Conference on Information Technology, Coding and Computing*, volume 1, pages 729–733, Las Vegas, Nevada, April 2004.

Joel J. Mintzes, James H. Wandersee, and Joseph D. Novak. *Assessing science understanding: A human constructivist view.* Academic Press, 2005.

Dan I. Moldovan, Christine Clark, and Moldovan Bowden. Lymba's PowerAnswer 4 in TREC 2007. In *TREC*, 2007.

Jack Mostow and Wei Chen. Generating instruction automatically for the reading strategy of self-questioning. In Vania Dimitrova, Riichiro Mizoguchi, Benedict du Boulay, and Art Graesser, editors, *Proceeding of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, pages 465–472, Amsterdam, The Netherlands, 2009. IOS Press.

Tom Murray. Authoring Knowledge-Based tutors: Tools for content, instructional strategy, student model, and interface design. *Journal of the Learning Sciences*, 7(1):5, 1998. ISSN 1050-8406.

Roberto Navigli and Paola Velardi. From glossaries to ontologies: Extracting semantic structure from textual definitions. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 71–87, Amsterdam, The Netherlands, The Netherlands, 2008. IOS Press. ISBN 978-1-58603-818-2.

Victor U. Odafe. Students generating test items: A teaching and assessment strategy. *Mathematics Teacher*, 91(3):198–203, 1998.

Andrew Olney, Whitney Cade, and Claire Williams. Generating concept map exercises from textbooks. In *Proceedings of the Sixth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 111–119, Portland, Oregon, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W11-1414`.

Andrew M. Olney. Extraction of concept maps from textbooks for domain modeling. In Vincent Aleven, Judy Kay, and Jack Mostow, editors, *Intelligent Tutoring Systems*, volume 6095 of *Lecture Notes in Computer Science*, pages 390–392. Springer Berlin / Heidelberg, 2010. URL `http://dx.doi.org/10.1007/978-3-642-13437-1\_80`. 10.1007/978-3-642-13437-1_80.

Andrew M. Olney, Arthur C. Graesser, and Natalie K. Person. Tutorial dialog in natural language. In R. Nkambou, J. Bourdeau, and R. Mizoguchi, editors, *Advances in Intelligent Tutoring Systems*, volume 308 of *Studies in Computational Intelligence*, pages 181–206. Springer-Verlag, Berlin, 2010.

Gary M. Olson, Susan A. Duffy, and Robert L. Mack. Question-asking as a component of text comprehension. In Arthur C. Graesser and J. B. Black, editors, *The psychology of questions*, pages 219–226. Lawrence Earlbaum, Hillsdale, N.J., 1985.

Jose Otero and Arthur C. Graesser. PREG: Elements of a model of question asking. *Cognition and Instruction*, 19(2):143–175, 2001.

Santanu Pal, Tapabrata Mondal, Partha Pakray, Dipankar Das, and Sivaji Bandyopadhyay. QG-STEC system description JUQGG: A rule based approach. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 76–79, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Annemarie S. Palinscar and Ann L. Brown. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1:117–175, 1984.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, 2005. ISSN 0891-2017. doi: http://dx.doi.org/10.1162/0891201053630264.

Natalie K. Person, Arthur C. Graesser, Joseph P. Magliano, and Roger J. Kreuz. Inferring what the student knows in one-to-one tutoring: the role of student questions and answers. *Learning and individual differences*, 6(2):205–229, 1994.

Jean Piaget. *The origins of intelligence*. International University Press, New York, 1952.

Michael Pressley and Donna Forrest-Pressley. Questions and children's cognitive processing. In A.C. Graesser and B. John, editors, *The Psychology of Questions*, pages 277–296. Lawrence Erlbaum Associates, Hillsdale, NJ, 1985.

Barak Rosenshine, Carla Meister, and Saul Chapman. Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, 66(2):181–221, 1996.

Vasile Rus and Arthur C. Graesser. The question generation shared task and evaluation challenge. Technical report, University of Memphis, 2009. ISBN:978-0-615-27428-7.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Svetlana Stoyanchev, and Cristian Moldovan. Overview of the first question generation shared task evaluation challenge. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 45–57, Pittsburgh, June 2010a. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Vasile Rus, Brendan Wyse, Paul Piwek, Mihai Lintean, Stoyanchev Svetlana, and Cristian Moldovan. The First Question Generation Shared Task Evaluation Challenge. In *Proceedings of the 6th International Natural Language Generation Conference (INLG 2010)*, Dublin, Ireland, 2010b.

Ivan A. Sag and Dan Flickinger. Generating questions with deep reversible grammars. In *Workshop on the Question Generation Shared Task and Evaluation Challenge*, Arlington, VA, September 2008.

Marlene Scardamalia and Carl Bereiter. Fostering the development of self-regulation in children's knowledge processing. In Susan F. Chipman, Judity W. Segal, and Robert Glaser, editors, *Thinking and learning skills*, volume 2, pages 563–577. Erlbaum, Hillsdale, NJ, 1985.

Roger C. Schank. *Dynamic memory revisited*. Cambridge University Press, 1999. ISBN 0521633982, 9780521633987.

Harry Singer and Dan Donlan. Active comprehension: Problem-solving schema with question generation for comprehension of complex short stories. *Reading Research Quarterly*, 17(2):166–186, 1982.

John F. Sowa. Semantic networks. In Stuart C Shapiro, editor, *Encyclopedia of Artificial Intelligence*. Wiley, 1992.

Robert D. Tennyson and Ok-Choon Park. The teaching of concepts: A review of instructional design research literature. *Review of Educational Research*, 50(1):55 –70, 1980. doi: 10.3102/00346543050001055. URL `http://rer.sagepub.com/content/50/1/55.abstract`.

Alejandro Valerio and David B. Leake. Associating documents to concept maps in context. In A. J. Canas, P. Reiska, M. Ahlberg, and J. D. Novak, editors, *Proceedings of the Third International Conference on Concept Mapping*, 2008.

Lucy Vanderwende. The importance of being important: Question generation. In Vasile Rus and Art Graesser, editors, *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, September 25-26 2008.

Andrea Varga and Le An Ha. Wlv: A question generation system for the qgstec 2010 task b. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 80–83, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Ellen M. Voorhees and Hoa Trang Dang. Overview of the TREC 2005 question answering track. In *NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2005. URL `citeseer.ist.psu.edu/article/voorhees02overview.html`.

Terry Winograd. *Understanding Natural Language*. Academic Press, New York, 1972.

Bernice Y. L. Wong. Self-questioning instructional research: A review. *Review of Educational Research*, 55(2):227–268, 1985.

Brendan Wyse and Paul Piwek. Generating questions from OpenLearn study units. In Scotty D. Craig and Darina Dicheva, editors, *The 2nd Workshop on Question Generation*, volume 1 of *AIED 2009 Workshops Proceedings*, pages 66–73, Brighton, UK, July 2009. International Artificial Intelligence in Education Society.

Xuchen Yao. Question generation with minimal recursion semantics. Master's thesis, Saarland University & University of Groningen, 2010. URL `http://cs.jhu.edu/~xuchen/paper/Yao2010Master.pdf`.

Xuchen Yao and Yi Zhang. Question generation with minimal recursion semantics. In Kristy Elizabeth Boyer and Paul Piwek, editors, *Proceedings of QG2010: The Third Workshop on Question Generation*, pages 68–75, Pittsburgh, June 2010. questiongeneration.org. URL `http://oro.open.ac.uk/22343/`.

Barry J. Zimmerman. A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3):329–339, 1989.

Amal Zouaq and Roger Nkambou. Evaluating the generation of domain ontologies in the knowledge puzzle project. *IEEE Trans. on Knowl. and Data Eng.*, 21(11):1559–1572, 2009. ISSN 1041-4347. doi: http://dx.doi.org/10.1109/TKDE.2009.25.