Automatically Generating Questions about Novel Metaphors in Literature

Natalie Parde
Computer Science
University of Illinois at Chicago
parde@uic.edu

Rodney D. Nielsen
Computer Science and Engineering
University of North Texas
rodney.nielsen@unt.edu

Abstract

The automatic generation of stimulating questions is crucial to the development of intelligent cognitive exercise applications. We developed an approach that generates appropriate *Questioning the Author* queries based on novel metaphors in diverse syntactic relations in literature. We show that the generated questions are comparable to human-generated questions in terms of naturalness, sensibility, and depth, and score slightly higher than human-generated questions in terms of clarity. We also show that questions generated about novel metaphors are rated as cognitively deeper than questions generated about non- or conventional metaphors, providing evidence that metaphor novelty can be leveraged to promote cognitive exercise.

1 Introduction

Automatic question generation is useful for a wide range of applications, including those providing educational and cognitive exercise support. Most question generation work to date has focused on generating *factoid* questions—that is, questions regarding factual content that is readily available in the source text. Factoid questions are well-suited to some contexts, such as quizzing for simple comprehension. However, answering them typically requires only shallow reasoning skills, rendering them unsuitable for situations in which deeper cognitive engagement is desired.

Less work has been conducted with the goal of generating deeper questions. We work toward filling that void by presenting an approach for automatically generating questions about *novel metaphors* from popular classic fiction using the *Questioning the Author* strategy. Metaphor novelty is defined here as the degree of likelihood with which one can expect to encounter a given metaphor on a regular basis. Consider the sentence:

> I *spent* an *hour* on my homework.

The word pair, {*spent*, *hour*}, is a highly conventional metaphor—although one cannot literally spend time, phrases such as this are highly common in the English language. Alternately, consider:

> The Queen was *frowning* like a *thunderstorm*.

The word pair, {*frowning*, *thunderstorm*}, is a highly novel metaphor. Novel metaphors reside at the opposite end of the continuum from conventional metaphors, and should strike one as being particularly interesting or creative. These are the metaphors of particular interest to us in this work.

The targeted focus on novel metaphors stems from prior work showing that novel metaphors are more difficult to process, both in young adults (Lai et al., 2009) and in older adults with and without dementia (Amanzio et al., 2008; Mashal et al., 2011). The latter are a key demographic for our use case, an elder-focused human-robot book discussion system (Parde, 2018). Here, we (1) introduce a method for generating deep questions about diverse novel metaphors following the Questioning the Author strategy. We (2) show that the resulting questions are comparable to or score slightly higher than questions generated by everyday users about the same topics in terms of naturalness, clarity, sensibility, and depth. Moreover, we (3) provide empirical evidence that questions automatically generated about novel metaphors are rated as having greater depth than questions

Proceedings of The 11th International Natural Language Generation Conference, pages 264–273,
Tilburg, The Netherlands, November 5-8, 2018.

automatically generated about non- or conventional metaphors. Finally, we (4) publicly release our source code and a corpus of question ratings and responses for both human- and automatically-generated questions to the research community to foster additional work in this area.

2 Related Work

Automatic question generation and its potential to facilitate learning has been of interest to researchers since at least the 1970s, when John Wolfe introduced the pattern-matching AUTOQUEST question generation algorithm (Wolfe, 1977). Today, many question generation systems exist for educational applications, with most generating factoid questions about content found in expository text (Araki et al., 2016; Du et al., 2017; Gates, 2008; Heilman and Smith, 2010; Mazidi and Nielsen, 2014, 2015; Rus et al., 2007; Serban et al., 2016; Wyse and Piwek, 2009).

These systems achieve their goals in a number of ways, including templates (Araki et al., 2016; Mazidi and Nielsen, 2014, 2015; Rus et al., 2007; Wyse and Piwek, 2009), sentence transformations (Gates, 2008; Heilman and Smith, 2010), and recently, neural networks (Du et al., 2017; Serban et al., 2016). Template-based systems select templates based on syntactic structure, semantic role labels, dependency parses, and/or discourse cues to produce generally shallower questions (e.g., "Inflation is defined as an increase in the price level." → "How is inflation defined?" (Mazidi and Nielsen, 2015)). The template-based system developed by Araki et al. (2016) generated questions over multiple sentences to produce questions that required more inference steps than those generated from a single sentence, using event coreference, entity coreference, and paraphrases. However, the answers to these questions were still readily available in the original text passage (in fact, ensuring that this was the case was a goal of the system). Although shallower questions are suitable for quizzing comprehension of expository text (the most common scenario to which they are applied), they are inadequate for more involved discussions, such as those analyzing fiction narrative.

Deeper questions (or more aptly, writing prompts) were generated by Liu et al.'s (2012) system, designed to help students write better literature reviews. Sentences containing citations were classified as describing opinions, methods, results, or one of several other categories, and templates were selected based on those classifications to construct questions using content from the original sentence (e.g., "Cannon (1927) challenged this view mentioning that physiological changes were not sufficient to discriminate emotions." → "Why did Cannon challenge this view mentioning that physiological changes were not sufficient to discriminate emotions? (What evidence is provided by Cannon to prove the opinion?) Does any other scholar agree or disagree with Cannon?"). Lindberg et al.'s (2013) system prompted students for summaries, causal effects, and descriptions not expected to be answerable from the immediate sentence from which they were generated, using question templates selected based on semantic role patterns. The questions were then classified as having/not having learning value, allowing the system to automatically discard poor-quality questions. The learning value classifier was trained using length, language model, semantic role label, named entity, glossary, and syntax features.

The system developed by Becker et al. (2012) automatically identified question topics (i.e., the part of a sentence about which a question should be asked), and accordingly generated cloze (fill-in-the-blank) questions. However, cloze questions are shallow and have limited potential to stimulate deep reasoning. Olney et al. (2012) developed a system that automatically extracted concept maps from expository text, and generated questions based on those concept maps using templates associated with different types of start nodes, end nodes, and edge relations from the maps. Finally, Mostow and Chen (2009) developed a method for automatically generating self-questioning instruction for students reading children's stories. Although some stages of this instruction were scripted, others involved automatically generating example questions for students. The template-based generation approach resulted in "why" questions about mental states expressed in the stories (e.g., "And when the country mouse saw the cheese, cake, honey, jam and other goodies at the house, he was pleasantly surprised." → "Why was the country mouse surprised?"). Although these approaches for generating deeper questions are promising, none have specifically sought to implement the Questioning the Author paradigm, which revolves around building meaning from text rather than quizzing a reader's com-

prehension. Moreover, none focus on generating questions based on identified occurrences of highly novel metaphor in the source text.

2.1 Questioning the Author

The questions generated for the work described herein employ a questioning strategy commonly used in K-12 education known as *Questioning the Author* (*QtA*). QtA seeks to encourage readers to consider the author's underlying intentions when crafting literary prose (Beck and McKeown, 2006). The strategy can be implemented with either expository text or fiction narrative. One of the key goals of QtA is to coax readers toward building meaning from and understanding the relationships between different elements or events present in the text, as opposed to focusing on isolated components as factoid questions are likely to do. Example QtA queries (Beck and McKeown, 2006) may include prompts such as:

- What is the author trying to say here?
- What do you think the author wants us to know?
- What is the author talking about?
- So what does the author mean right here?
- That's what the author said, but what did the author mean?

Such questions are open-ended, and typically elicit more detailed, free-form responses than factoid questions. They also typically encourage deeper analysis of the source text.

3 Template Development and Selection

We built templates based on sample questions from the book on QtA (Beck and McKeown, 2006), with slots to be filled using predicted novel metaphors that were automatically identified in literature. We describe the metaphor novelty scoring methodology in greater detail in Section 4.1. The identified metaphors were all syntactically-related pairs of words; we constructed different sets of templates for different syntactic relation types. The syntactic relation types for which questions could be generated included the *universal dependency relations* (McDonald et al., 2013) in Table 1. Some of the resulting templates are shown in Table 2; the full list of 130+ templates can be found online.[1]

Relation Type	Description	Example
nsubj	Nominal subject.	The *apple* is *red*.
nsubjpass	Nominal subject of a passive verb.	*Newton* was *hit* by an apple.
dobj	Direct object.	I *gave* him an *apple*.
iobj	Indirect object.	I *gave him* an apple.
csubj	Clausal subject.	What he *wanted* was an *apple*.
csubjpass	Clausal subject of a passive verb.	What he *wanted* as *taken* to be an apple.
xcomp	Open clausal component.	He *asked* to *eat* an apple.
nmod	Nominal modifier.	The *stem* of the *apple*.
acl	Clause that modifies a noun.	I need a *way* to *get* an apple.
appos	Appositional modifier.	The *fruit*, an *apple*, was red.
amod	Adjectival modifier.	It was a *red apple*.
advcl	Adverbial clause modifier.	He *got* is idea as the apple was *falling*.
dep	Dependency for which the parser cannot determine a finer-grained relation.	N/A
advmod	Adverbial modifier.	He *ate* the apple *quickly*.
compound	Multiword expression.	The apple had *polka dots*.

Table 1: Dependency relation types for which questions were generated.

Templates were chosen randomly from among the pool of all relevant templates for a given dependency type. For example, if a question was to be generated about a metaphor formed by two words syntactically related to one another using an *nsubj* dependency, a random selection from all possible templates corresponding to the *nsubj* type would be made. Surface realizations were then constructed by fitting predicted novel metaphors into the selected templates. That process is described in the following section.

4 Surface Realization

Realization was performed based on the linguistic characteristics and syntactic parse details corresponding to the novel metaphors about which questions were generated. Our procedure for iden-

[1] http://natalieparde.com/papers/inlg_question_templates.pdf

Dependency	Template
nsubj	What is the author trying to say with the expression '<DEP (N/J/P)> <GOV (V)>'?
nsubjpass	What do you think the author wants us to know by figuratively saying '<DEP (N/J/P)>' <WAS/WERE> '<GOV (V)>'?
dobj/iobj	What is the author talking about when <HE/SHE> writes '<GOV (V)>' <? THE> '<DEP (N/J/P)>'?
csubj	What's the important message in the expression '<CLAUSE>'?
csubjpass	So what does the author mean when <HE/SHE> writes '<CLAUSE_PASS>'?
xcomp/nmod/acl	The author said '<STRING>,' but what did <HE/SHE> mean?
appos	Does the expression '<DEP>' with '<GOV>' make sense?
amod	How does the expression '<DEP> <GOV>' fit in with what the author told us?
advcl/dep	Does the author tell us why <HE/SHE> wrote '<PHRASE>'?
advmod	Why do you think the author tells us '<DEP (V/J/R)> <GOV (V/J/R)>'?
compound	What is the author telling us with the expression '<W1> <W2>'?

Table 2: Example question templates.

tifying particularly novel metaphors (Parde and Nielsen, 2018b) and our methods for incorporating the approach in this work are described in Subsection 4.1. Subsection 4.2 describes how template slots were subsequently filled using the identified novel metaphors.

4.1 Metaphor Novelty Scoring

Our metaphor novelty scoring approach predicts continuous scores for syntactically-related pairs of content words (nouns, verbs, adjectives, and adverbs), with higher scores reflecting greater novelty than lower scores (Parde and Nielsen, 2018b). It consists of a four-layer feedforward neural network trained using features based on psycholinguistic characteristics (concreteness, imageability, sentiment, and ambiguity), word co-occurrence, syntactic structure, semantic characteristics, and information from WordNet (Miller, 1995) regarding the words in the pair. For the work here, we trained our neural network model on a corpus of word pairs originally extracted from the VU Amsterdam Metaphor Corpus (Steen et al., 2010) and labeled along a continuous scale for metaphor novelty (Parde and Nielsen, 2018a); the VU Amsterdam Metaphor Corpus is comprised of fiction, news articles, academic articles, and transcribed conversations. We then applied the learned model to all word pairs (52,279 total) extracted from a subset of sentences from 58 books that are publicly available on Project Gutenberg.[2] Finally, we randomly selected a small subset (457) of the word pairs having predicted scores greater than 1.0[3] as the identified novel metaphors about which to generate questions.

4.2 Slot Filling

As shown in Table 2, each template contains one or more slots: <GOV>, <DEP>, <WAS/WERE>, <HE/SHE>, <? THE>, <CLAUSE>, <CLAUSE_PASS>, <STRING>, <W1>, and <W2>. Filling the <GOV> and <DEP> slots is straightforward; the governor and modifier of the syntactic relation forming the predicted metaphor are merely substituted into the appropriate slots in the question template. The <HE/SHE> slot is the only slot requiring metadata about the source text being discussed (a gender was manually assigned to each book).

The token "was" or "were" is selected to fill the <WAS/WERE> slot based on the part-of-speech tags associated with the two words forming the predicted metaphor. Metaphors including plural nouns are given the verb "were," and all other metaphors are given the verb "was." The word "the" is optionally included in realizations of templates including the <? THE> slot based on the distance between the two words forming the predicted metaphor; if they are immediately next to one another, it is omitted, and otherwise it is included.

Filling the <CLAUSE> slot is more complex. A full dependency parse of the predicted metaphor's source sentence is first acquired using Stanford CoreNLP (Manning et al., 2014). A clause is then constructed using only tokens that are syntactically related to words forming the metaphor, in the order in which they occur in the source sentence. Consider the example sentence:

[2] https://www.gutenberg.org; the books selected included all books written in or translated to English and classified as fiction in the "Top 100 Books Over The Last 30 Days" list as of May 18, 2017.

[3] Across all word pairs, novelty predictions ranged from 0.24-1.41.

> What he *tasted* on this dark and stormy night was a *dream*.

To fit {*tasted*, *dream*} into the template, "What's the important message in the expression '<CLAUSE>'?", the following words would be identified as syntactically related to *tasted* and *dream*: {What, he, *tasted*, was, a, *dream*}. The realized question, retaining the words in their original order, would thus be: "What's the important message in the expression 'What he tasted was a dream'?" <CLAUSE_PASS> is constructed similarly, but requires only words syntactically related to the modifier.

The <STRING> slot was filled by tokenizing the source sentence, and extracting the full span of text from one of the words in the metaphor up to and including the other. Consider the sentence:

> She smelled a *melody* of *appetizers* and knew she had reached the networking event.

The word pair {*melody*, *appetizers*} would fit into the *nmod* template "The author said '<STRING>,' but what did <HE/SHE> mean?" as "The author said 'melody of appetizers,' but what did she mean?" The <PHRASE> slot was filled using a slightly broader window of text: the span reaching from the first word syntactically related to either of the words in the metaphor, to the last word syntactically related to either of those words, inclusive of the words forming the metaphor. Finally, <W1> and <W2> were filled simply by substituting the word from the metaphor that occurred first in the source sentence for <W1>, and the word that occurred second in the source sentence for <W2>.

5 Evaluation

The quality of the automatically-generated questions was evaluated relative to that of questions written by humans. The human-generated questions were comprised of two subsets: (1) those generated based on *sentences* containing predicted novel metaphors, and (2) those generated based on *pairs of words* predicted to be novel metaphors.

5.1 Data Collection

We crowdsourced human-generated questions based on a randomly-selected subset of the same 457 predicted novel metaphors about which questions were automatically generated, using Amazon Mechanical Turk (AMT).[4] Workers were simply instructed to create "good" questions, such as what they might ask in a book discussion group if they came across the sentence (or the bolded word pair within that sentence) when reading. These instructions were purposely open-ended to foster diversity in the collected data. To that end, we also continued to collect questions until the human-generated question dataset included 35 unique question authors (180 questions; 90 of each type). Sample questions collected based on sentences and word pairs are shown in Tables 3 and 4.

Source Sentence	Question Received
An icy horror of loneliness seized him; he saw himself standing apart and watching all the world fade away from him – a world of shadows, of fickle dreams.	What did he feel as he stood alone?
Perhaps, from the casement, standing hand-in-hand, they were watching the calm moonlight on the river, while from the distant halls the boisterous revelry floated in broken bursts of faint-heard din and tumult.	Do you think the people holding hands are supposed to be happy or sad?
I had crossed a marshy tract full of willows, bulrushes, and odd, outlandish, swampy trees; and I had now come out upon the skirts of an open piece of undulating, sandy country, about a mile long, dotted with a few pines and a great number of contorted trees, not unlike the oak in growth, but pale in the foliage, like willows.	Do you think the landscape reflects his inner feelings?
I quickly destroyed part of my sledge to construct oars, and by these means was enabled, with infinite fatigue, to move my ice raft in the direction of your ship.	If you were to make oars that way, how long do you think it would take?

Table 3: Sample questions generated by humans based on sentences.

The 180 human-generated and 457 automatically-generated questions were intermixed, and responses to the questions and ratings for four criteria for each question were also solicited using a separate pool of workers

[4] https://www.mturk.com; we crowdsourced questions from everyday users to facilitate comparison with the most likely alternative to a human-robot book discussion—a typical human book club.

Source Sentence	Question Received
Wavewhite wedded words shimmering on the **dim tide**.	what does dim tide mean?
All about me gathered the invisible terrors of the Martians; that **pitiless sword** of heat seemed whirling to and fro, flourishing overhead before it descended and smote me out of life.	Why would the martians kill him?
My father saw this change with pleasure, and he turned his thoughts towards the best method of eradicating the remains of my melancholy, which every now and then would return by fits, and with a **devouring blackness** overcast the approaching sunshine.	How is the image of mortality described with the weather?
But the **overflowing misery** I now felt, and the excess of agitation that I endured rendered me incapable of any exertion.	How did things get so bad that they essentially felt overflowing misery?

Table 4: Sample questions generated generated by humans based on word pairs.

from AMT. The criteria considered were as follows:

- **Naturalness:** The degree to which the question seems natural, or sounds "normal" to the reader.
- **Clarity:** The degree to which it is clear to the reader how he or she is supposed to respond to the question, regardless of whether he or she is sure of the answer.
- **Sensibility:** The degree to which the reader feels it makes sense to ask the question, given the source sentence upon which it is based.
- **Depth:** The degree to which the reader feels challenged in coming up with an answer to the question.

Two workers rated each criterion using a five-point scale. Small disagreements were adjudicated by averaging, and disagreements greater than a difference of 2.0 (e.g., a 1 and a 4) were forwarded to a third-party, native English speaking adjudicator (211 questions required adjudication for at least one of the four criteria). Overall, the crowd workers exhibited moderate agreement with one another, with Krippendorff's α=0.50, α=0.52, α=0.51, and α=0.52 for ratings of naturalness, clarity, sensibility, and depth, respectively. The collected question answers are not used in this work, but they serve the plural purpose of making the dataset more broadly useful, lending insight regarding the types of answers expected to inform future work on question generation and response scaffolding, and providing a coarse-grained quantitative (time-based) measure of question depth.

5.2 Average Question Ratings

Average ratings for the question criteria, both overall and when only considering questions for a given criterion that had received above-midpoint (> 3.0) ratings for the previous criteria, are presented in Tables 5 and 6. The latter scenario was included to reduce the potential for confusion in interpreting the results (for instance, unclear questions that were also unnatural may have only been rated as such because they were unnatural; these questions are included in the average score reported in Table 5 but not in the average score reported in Table 6). To elaborate further, the constraints considered in the latter scenario (as well as for the results reported in Tables 7 and 8) were:

- Naturalness: All ratings were considered.
- Clarity: Only questions having a *Naturalness* score > 3.0 were considered.
- Sensibility: Only questions having *Naturalness* and *Clarity* scores > 3.0 were considered.
- Depth: Only questions having *Naturalness*, *Clarity*, and *Sensibility* scores > 3.0 were considered.

Significance values for both scenarios were determined via one-way ANOVA between the two groups. Not surprisingly, given the instructions to ask *good* questions, automatically-generated questions did not quite match the high bar set for depth by humans' questions, but this difference was not statistically significant. The only significant ($p < 0.05$) difference reported between the two groups in Table 6 was for ratings of *Clarity* (automatically-generated questions scored slightly higher). This finding was echoed when considering the overall averages (Table 5); again, the only statistically significant difference between groups was that ratings of *Clarity* were slightly higher for the automatically-generated questions than the human-generated questions.

	Human-Generated	Automatically-Generated	*p*
Naturalness	3.89	4.02	0.13
Clarity	3.78	4.04	0.00
Sensibility	3.83	4.00	0.06
Depth	3.78	3.67	0.24

Table 5: Average ratings across all question criteria, with significance values.

	Human-Generated	Automatically-Generated	*p*
Naturalness	3.89	4.02	0.13
Clarity	4.18	4.34	0.03
Sensibility	4.45	4.48	0.61
Depth	3.92	3.76	0.17

Table 6: Average ratings, considering only questions with above-average ratings for the preceding criteria.

	Human-Generated (Sentence)	Human-Generated (Word Pair)	Auto.-Generated
Nat.	3.92	3.85	4.02
Clar.	4.08	4.29	4.34
Sens.	4.30	4.42	4.48
Depth	3.91	4.03	3.76

Table 7: Average ratings for human-generated question subgroups and automatically-generated questions.

	TP	FP	*p*
Naturalness	3.99	4.06	0.42
Clarity	4.23	4.46	0.00
Sensibility	4.31	4.44	0.09
Depth	3.92	3.64	0.02

Table 8: Average ratings for true and false positives among automatically-generated questions.

5.3 Average Ratings for Question Subgroups

In addition to these broad comparisons of human- and automatically-generated questions, we examined the differences between different subgroups. Table 7 presents the average ratings for (1) human-generated questions based on sentences, (2) human-generated questions based on specified word pairs, and (3) automatically-generated questions. Statistical significance was computed using one-way ANOVAs beween each pair of groups: human-generated (sentence) and human-generated (word pair); human-generated (sentence) and automatically-generated; and human-generated (word pair) and automatically-generated. Only two statistically significant differences existed between the subgroups: the average ratings for *Clarity* and *Sensibility* were higher for automatically-generated questions than for human-generated questions based on sentences. These differences were not statistically significant when comparing human-generated questions based on word pairs and automatically-generated questions.

Table 8 presents average ratings for two subsets of automatically-generated questions: true positives (TP) for which the word pair about which the question was generated was both predicted to be a novel metaphor and actually was a novel metaphor, and false positives (FP) for which the word pair about which the question was generated was predicted to be a novel metaphor but was not actually a novel metaphor. We collected gold standard metaphor novelty scores for these word pairs in the same manner by which we built our previous VUAMC-based metaphor novelty dataset (Parde and Nielsen, 2018a), used to train the metaphor novelty prediction model in this work. Specifically, we crowdsourced five annotations for each word pair, and automatically aggregated them to continuous scores using a label aggregation model learned from features based on annotation distribution and presumed worker trustworthiness (Parde and Nielsen, 2017).

There were two statistically significant differences between the two groups: questions about false positives were rated as clearer than questions about true positives, and questions about true positives were rated as having more depth than questions about false positives. One hypothesis regarding the former finding is simply that non-metaphoric language is more clearly interpretable than metaphoric language. The finding that question depth is higher for questions about true positives (novel metaphors) than questions about other instances provides empirical support for the underlying motivations guiding this work—namely, that questions regarding novel metaphors are more cognitively challenging than similar questions about non-metaphors or conventional metaphors.

5.4 Correlations between Question Criteria

In addition to evaluating question quality on the basis of average ratings for each question criterion, we computed Pearson's correlation scores

	Nat.	Clar.	Sens.	Depth	Compl. Time
Nat.	-	0.55	0.48	0.03	-0.04
Clar.		-	0.65	0.05	-0.01
Sens.			-	0.05	-0.04
Depth				-	0.07
Compl. Time					-

Table 9: Correlations between categories of ratings for both automatically- and human-generated questions.

between the four criteria, as well as between those criteria and completion time (the amount of time workers took to complete each HIT, including rating the four criteria and writing a response to the question, on AMT) to examine which of these factors were correlated with one another. Questions rated as being natural, clear, and sensible (scores > 3.0) were included in this evaluation. Since a small number of HITs had outlier completion times far exceeding the average (indicating that the rater most likely left their browser open while taking a break, rather than actually spending that much time completing the HIT itself), we removed HITs with completion times +/- two standard deviations from the mean completion time from consideration.

Table 9 presents a matrix of overall correlation scores when considering ratings for both automatically-generated and human-generated questions. Overall, moderately strong positive correlations were found between naturalness and clarity (r=0.55), naturalness and sensibility (r=0.48), and clarity and sensibility (r=0.65). No strong correlations were found between these criteria and depth or completion time. When computing correlations only between ratings collected for human-generated questions or for only automatically-generated questions, these trends persisted. In addition to the correlations reported in Table 9, we computed the correlation between completion time and *metaphor novelty* for each set, finding correlations of r=0.09 across all questions, r=0.19 for human-generated questions, and r=0.03 for automatically-generated questions.

5.5 Discussion and Future Recommendations

The findings regarding the average ratings overall and for different subgroups, as well as regarding the correlations between types of ratings, provide interesting and in a few cases (such as the higher average *Clarity* score for automatically-generated questions rather than human-generated questions) slightly surprising observations. It is clear that the automatically-generated questions are very comparable with human-generated questions in terms of all criteria considered. **Across all comparisons, there were no cases in which an average rating associated with human-generated questions or a subset thereof statistically significantly outperformed an average rating for the same criterion with automatically-generated questions.** As such, it is reasonable to assume that the approach is capable of generating sufficiently natural, clear, sensible, and challenging questions relative to what the average person might generate.

That said, there are still some areas that could be improved upon. For example, the average depth rating for automatically-generated questions that were also rated as natural, clear, and sensible (as measured by having ratings greater than 3.0) was 3.76. Although this is above mid-range, it could certainly be higher (the maximum score allowed was a 5.0). Thus, additional work could be done to improve upon question depth in future work. This could perhaps be accomplished by introducing complementary strategies to QtA. Considerations could also be taken to identify optimal questioning *sequences*—that is, algorithmically deciding upon groups of questions most likely to challenge readers when asked in sequence, as opposed to simply selecting questions at the individual level.

Future work toward improved metaphor novelty scoring algorithms will result in a higher likelihood that the subjects of the automatically-generated questions are indeed novel metaphors. The evaluation indicates that improved identification of novel metaphors should lead to higher average ratings of question depth. Specifically, Table 8 shows that in a comparison of automatically-generated questions for true positives (instances predicted to be novel metaphors that were actually novel metaphors) versus false positives (instances predicted to be novel metaphors that were actually not), **questions generated for novel metaphors were rated as having more depth than similar questions generated for conventional or non-metaphors, and this difference was found to be statistically significant**.

Finally, many of the correlation scores observed between rating criteria were expected (it is difficult to think of questions that are, for instance, highly

natural-sounding while also unclear). We had anticipated a slightly higher positive correlation than was observed between question depth and completion time, as a natural assumption is that if a question makes a reader think quite a bit before answering it, it will take longer to formulate an answer than if the question doesn't make a reader think at all. However, the measurement of completion time was coarse-grained; it only considered the overall amount of time that it took the worker to complete the full HIT including reading the instructions and a source sentence, rating four criteria, and finally constructing a written response to the question. Many variables outside of the question depth itself could therefore impact the overall completion time. In the future, work can be conducted to examine the correlation between completion time and question depth in a more controlled environment.

6 Conclusion

In this work, we introduced and evaluated a question generation approach to automatically construct QtA queries about predicted novel metaphors. We designed and validated question templates based on sample questions drawn directly from the book on QtA (Beck and McKeown, 2006), and demonstrated methods capable of producing high-quality question realizations. We evaluated the automatically-generated questions relative to human-generated questions based on the same source material, and discovered that the only statistically significant difference between the two groups with respect to four distinct criteria (naturalness, clarity, sensibility, and depth) was that the automatically-generated questions received slightly higher clarity scores. We analyzed the correlations among the four question criteria as well as between the question criteria and completion time, and found strong positive correlations between naturalness, clarity, and sensibility, but only weak correlations between each of those criteria and question depth.

All data and source code are publicly available. Ultimately, our evaluation proved that questions about novel metaphors in literature can be automatically generated at a quality level comparable to what the average human might generate. It also provided empirical support for an underlying motivation guiding this work: that questions about novel metaphors can be leveraged as a means for motivating cognitive exercise.

Acknowledgements

We thank the anonymous reviewers for their comments and suggestions. This material was based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant 1144248, and the National Science Foundation under Grant 1262860. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Martina Amanzio, Giuliano Geminiani, Daniela Leotta, and Stefano Cappa. 2008. Metaphor comprehension in alzheimers disease: Novelty matters. *Brain and Language*, 107(1):1 – 10.

Jun Araki, Dheeraj Rajagopal, Sreecharan Sankaranarayanan, Susan Holm, Yukari Yamakawa, and Teruko Mitamura. 2016. Generating questions and multiple-choice answers using semantic analysis of texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1125–1136, Osaka, Japan. The COLING 2016 Organizing Committee.

Isabel L. Beck and Margaret G. McKeown. 2006. *Improving Comprehension with Questioning the Author: A Fresh and Expanded View of a Powerful Approach*. Theory and Practice. Scholastic.

Lee Becker, Sumit Basu, and Lucy Vanderwende. 2012. Mind the gap: Learning to choose gaps for question generation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 742–751, Montréal, Canada. Association for Computational Linguistics.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Donna M. Gates. 2008. Generating look-back strategy questions from expository texts. In *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, Los Angeles, California. Association for Computational Linguistics.

Vicky Tzuyin Lai, Tim Curran, and Lise Menn. 2009. Comprehending conventional and novel metaphors: An ERP study. *Brain Research*, 1284:145 – 155.

David Lindberg, Fred Popowich, John Nesbit, and Phil Winne. 2013. Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 105–114, Sofia, Bulgaria. Association for Computational Linguistics.

Ming Liu, Rafael A Calvo, and Vasile Rus. 2012. G-asks: An intelligent automatic question generation system for academic writing support. *Dialogue & Discourse*, 3(2):101–124.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Nira Mashal, Ronit Gavrieli, and Gitit Kav. 2011. Age-related changes in the appreciation of novel metaphoric semantic relations. *Aging, Neuropsychology, and Cognition*, 18(5):527–543. PMID: 21819177.

Karen Mazidi and Rodney D. Nielsen. 2014. *Pedagogical Evaluation of Automatically Generated Questions*. Springer International Publishing.

Karen Mazidi and Rodney D. Nielsen. 2015. *Leveraging Multiple Views of Text for Automatic Question Generation*. Springer International Publishing.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 92–97.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Jack Mostow and Wei Chen. 2009. Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings 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, The Netherlands. IOS Press.

Andrew M. Olney, Arthur C. Graesser, and Natalie K. Person. 2012. Question generation from concept maps. *Dialogue & Discourse*, 3(2):75–99.

Natalie Parde. 2018. Reading with robots: Towards a human-robot book discussion system for elderly adults. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18) Doctoral Consortium*, New Orleans, Louisiana. Association for the Advancement of Artificial Intelligence.

Natalie Parde and Rodney D. Nielsen. 2017. Finding patterns in noisy crowds: Regression-based annotation aggregation for crowdsourced data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1908–1913, Copenhagen, Denmark. Association for Computational Linguistics.

Natalie Parde and Rodney D. Nielsen. 2018a. A corpus of metaphor novelty scores for syntactically-related word pairs. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan. European Language Resources Association.

Natalie Parde and Rodney D. Nielsen. 2018b. Exploring the terrain of metaphor novelty: A regression-based approach for automatically scoring metaphors. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, New Orleans, Louisiana. Association for the Advancement of Artificial Intelligence.

Vasile Rus, Zhiqiang Cai, and Arthur C. Graesser. 2007. Experiments on generating questions about facts. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '07, pages 444–455, Berlin, Heidelberg. Springer-Verlag.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

John H. Wolfe. 1977. Reading retention as a function of method for generating interspersed questions. Technical report, San Diego: Navy Personnel Research and Development Center.

Brendan Wyse and Paul Piwek. 2009. Generating questions from openlearn study units. In *Proceedings of the AIED 2nd Workshop on Question Generation*, pages 66–73.