Towards Difficulty Controllable Selection of Next-Sentence Prediction Questions

Jingrong Feng
Language Technologies Institute
Carnegie Mellon University
jingronf@cs.cmu.edu

Jack Mostow
Robotics Institute
Carnegie Mellon University
mostow@cs.cmu.edu

ABSTRACT

Automatic Question Generation seeks to generate questions about a given text for educational purposes such as testing students' comprehension processes while reading. This paper focuses on the task of predicting the next sentence as a way to exercise and assess a crucial skill that comprehension questions often fail to test, namely relating sentences to the context preceding them. We train a BERT-based model of text coherence to estimate the probability that a given sentence will come next in a story. It achieves 68.4% AUC on a held-out test set, significantly above chance. We define an easiness score as the difference between the estimated probabilities of the next sentence and (the likelier of) two distractors, namely the two subsequent sentences. We evaluate our model on data from Project LISTEN's Reading Tutor by correlating the easiness scores of 1,023 questions against the percentage answered correctly by 274 children. A strong correlation would make it possible to filter such questions by difficulty for children at a specified reading level. Unfortunately, the easiness scores of the questions did not correlate with the correctness of children's answers to them.

Keywords

Automatic question generation, difficulty prediction, next-sentence prediction, reading comprehension assessment, natural language processing, BERT

1. INTRODUCTION

A crucial skill in reading comprehension is inter-sentential processing – integrating meaning across sentences. It involves analysis of cohesive relationships such as coreference, indirect reference, and ellipsis [3]. Inter-sentential processing is hard for young readers partly because it requires assimilation from short-term memory to mid-term memory [12]. Unfortunately, reading comprehension questions often fail to assess inter-sentential information integration [1, 13, 14].

Next-sentence prediction questions are a natural way to test inter-sentential processing and are easy to generate. They are also easy to score, because by definition the correct answer is the next sentence. One form of such questions is true/false, i.e., "Does this sentence come next?" Another form is multiple choice, i.e., "Which sentence comes next?" This form has a higher cognitive load because it requires considering multiple sentences, but may be easier than judging a single candidate sentence by itself. Figure 1 shows both.

Context: Everyone knows that the elephant has a very long nose. But a long time ago, the elephant's nose was short and fat. Like a shoe in the middle of its face.

Does this sentence come next?
− She had a question for every animal.

Which sentence comes next?
− She was curious about everything.
+ One day a baby elephant was born.
− She had a question for every animal.

Figure 1: Two forms of next-sentence prediction questions. Answers in green are correct and answers in red are incorrect.

Although easy to generate and score, next-sentence prediction questions can be hard to answer correctly. For example, one study [2] randomly inserted "Which sentence comes next?" questions in children's stories, with the next three sentences of the story in random order as the choices. Children answered only 41% of these questions correctly, barely above chance and frustratingly low.

Good questions should be challenging but not frustratingly hard. Therefore, difficulty control is important in automatic question generation. However, despite the rapid development of question generation, little work has analyzed the difficulty of automatically generated questions [9], especially for reading comprehension [6, 7, 16], and none of it addresses next-sentence prediction questions.

This paper addresses the difficulty of such questions, and is organized as follows. Section 2 describes how we trained a coherence model to estimate the probability that a sentence comes next given the preceding context, and how we used it to score question easiness. Section 3 evaluates this model on a corpus of children's stories. Section 4 correlates the easiness scores of the questions against the percentage of children who answered the questions correctly. Section 5 concludes.

2. COHERENCE ESTIMATION

To estimate the coherence between a given context and sentence, we fine-tuned a BERT-based binary classification model.

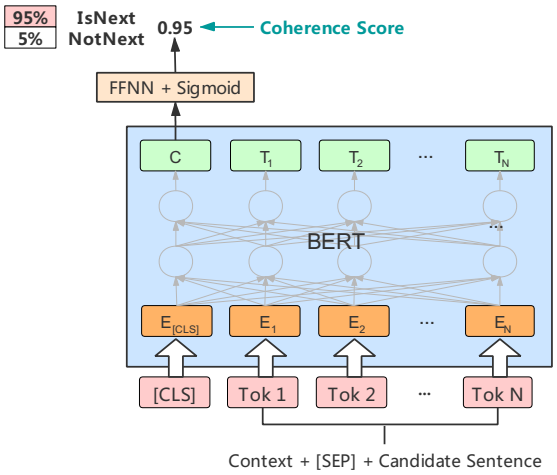

Figure 2: Architecture of the BERT-based model for coherence estimation.

BERT [5], a widely used Transformer-based language model, has achieved state-of-the-art performance on a large suite of natural language processing tasks. The blue box in Figure 2 shows the architecture of the pre-trained BERT model. To do classification, it appends a 2-layer feed-forward neural network (FFNN) to the BERT model, followed by a sigmoid function to scale the FFNN's output between 0 and 1.

BERT was pre-trained on BooksCorpus (800M words) [17] and English Wikipedia (2,500M words) with two objectives. First, randomly masking various words in a text and predicting the masked words from the surrounding text forced BERT to embed each word based on the surrounding words. Second, predicting whether one sentence follows another sentence in the original text forced BERT to learn inter-sentential coherence. Thus these two objectives prompted BERT to learn both intra- and inter-sentential semantic structure.

The effect of the next-sentence prediction task in pre-training has recently been questioned [4, 8, 15]. Some researchers believe that BERT actually learns inter-sentential topic similarity rather than coherence, because its negative instances are sentences sampled randomly from the entire text corpus, which are likely to be topically unrelated to the context.

We now describe how we adapted the BERT-based model to estimate inter-sentential coherence in children's stories.

Input: We fine-tuned the pre-trained BERT-based model on input token sequences of the following form:

- a special token `[CLS]` used for classification tasks
- three sentences of context, which we assume suffice to capture the semantically relevant content. Any more might include irrelevant information or exceed BERT's input length limit of 512 word pieces (i.e., roots and morphemes).
- a special separator token `[SEP]`
- a candidate next sentence; for positive instances, the sentence immediately following the context.

Selection of negative instances: We wanted the task to test children's judgment of inter-sentential coherence, not merely topical relevance. Therefore, rather than sample negative instances randomly from the entire corpus, we selected them from the same story, specifically the 2 sentences immediately following the correct sentence, which are likelier to be topically relevant to the local context than sentences from later in the story. Using the 3 sentences following the context as the multiple choice candidates also matched the task performed by the children in our evaluation dataset, to be described in Section 4.

Human experts could presumably pick contexts and distractors more judiciously to test children's judgements of inter-sentential coherence. However, such manual selection is neither economical nor scalable. One goal of this work was to identify requirements for choosing better contexts and distractors so as to improve automated selection.

Positive-negative ratio of training instances: BERT was pre-trained on equal numbers of positive and negative instances. In contrast, the 3 candidate sentences after the 3-sentence context included one positive instance and two negative instances.

Training labels: To fine-tune BERT and train the FFNN, we set the output of the combined model to 1 for positive instances and 0 for negative instances.

Easiness scores: To measure each candidate sentence's coherence with the given context, we used the probability output by the sigmoid function. Given this measure of coherence, we used a simple heuristic to rate the easiness e of answering a 3-choice question:

$$e = c_{pos} - max(c_{neg_1}, c_{neg_2}) \quad (1)$$

Here c_{pos} is the coherence of the correct answer, and c_{neg_i} is the coherence of distractor neg_i. This formula assumes that the difficulty of the question depends on whichever distractor is more coherent with the context. (As a reviewer suggested, we also tried the log ratio of the two coherence scores instead of their difference, but it performed the same in the evaluation reported in Section 4.)

Figures 4 and 5 show example questions with easiness scores of 0.244 and -0.262, respectively (see Appendix). A negative easiness score occurs when a distractor has greater coherence than the correct answer.

3. EVALUATION OF COHERENCE MODEL

We now evaluate how accurately our coherence model classified the 3 sentences following a 3-sentence context as $IsNext$ or $NotNext$.

3.1 Text Dataset

We constructed a dataset for fine-tuning and evaluating our coherence model from a corpus of English-language children's stories from two sources:

Table 1: Examples of Cases Removed by Data Cleaning

Type	Context	Correct Answer	Choices
two identical choices	...Did the frog slip?	Yes.	<Yes.> <Yes.> <The frog swam fast.>
one choice appearing in the context	...Yes.	Yes.	<Yes.> <The frog swam fast.> <It went past Pat.>
very short context	Pop can twist and bend. Pop slips! Pop stops.	Pop sits.	<Pop sits.> <Pat slaps Pop's hand.> <Pop must rub his feet!>
unfinished sentence in the context/content about phonics instruction	real meat peak	What sound do the letters e a make in the words real, meat, and peak?	<What sound do the letters e a make in the words real, meat, and peak?> <near> <leap>

- 337 stories from Project LISTEN's Reading Tutor [2], totalling 39K words with a vocabulary of 8K distinct words, at grade levels K-7.
- 354 stories from www.africanstorybook.org totalling 91K words with a vocabulary size of 11K, with page lengths ranging from one word to multiple paragraphs.

For fine-tuning and evaluation, we split the 337 LISTEN stories into three subsets, with 60% for training, 20% for hyper-parameter tuning, and 20% for testing, so as to ensure that stories in the test set were not seen during training. We used the African Storybook stories to augment the training set.

For every story in the corpus, we used a 6-sentence sliding window to generate next-sentence prediction items of the form ([`3-sentence context`; `IsNext sentence`; `NotNext sentence`; `NotNext sentence`]), with the correct (IsNext) sentence and two (NotNext) distractors to be presented in random order.

To clean the data, we filtered out several cases (illustrated in Table 1):

- **Cases with two identical choices or a choice appearing in the context:** typically caused by repeated sentences in a conversation.
- **Cases with context or a candidate sentence exceeding 125 words:** might cause the input sequence to exceed BERT's input length limit of 512 word pieces.
- **Cases with very short context:** typically caused by short sentences in a conversation that provide too little information to predict which sentence belongs next.
- **Cases with an unfinished sentence in the context:** for some poems or phonics instructions, sentences were not segmented according to sentence separators.
- **Cases about pronunciation or spelling:** are not relevant to semantic coherence.
- **Cases with the same context followed by different sentences:** may confuse the model during training.

As a result, we got a dataset consisting of 10,761 instances for training, a development set of 1,716 instances for hyper-parameter tuning, and a test set of 2,340 instances for evaluation.

3.2 Training

To fine-tune our coherence model, we used BERT_{base} [5] as the backbone, and the AdamW optimizer [10] with a initial learning rate of 1e-3 and a ReduceLROnPlateau scheduler[1]. We used a ReLU [11] activation in the hidden layer of the FFNN, and set the dropout probability of this hidden layer to 0.5. We trained the model with a standard binary cross-entropy loss function weighted by the positive-negative sample ratio of 1:2.

In contrast to pre-training BERT's hundreds of millions of parameters from scratch, fine-tuning the BERT-based coherence model was inexpensive. It took only about 5 minutes on a single Tesla-V100 GPU to optimize the parameters on the training set.

3.3 Evaluation Results

Table 2 evaluates the coherence model on the development and test sets using various metrics: accuracy, weighted-average precision, recall and F1-score, and area under the ROC curve (AUC). To evaluate metrics other than AUC, we set the classification threshold to 0.5 and compared the predicted label with the ground truth label. In other words, we classified an instance as $IsNext$ if the output probability (coherence score) exceeded this threshold, otherwise as $NotNext$. AUC measures the entire area beneath the ROC curve, which plots true positive rate vs. false positive rate at different classification thresholds. AUC evaluates the overall performance of a classification model by aggregating across all possible classification thresholds.

Table 2: Evaluation of the Coherence Model

Dataset	Accuracy	Precision	Recall	F1-score	AUC
Dev	0.608	0.663	0.608	0.620	0.662
Test	0.609	0.679	0.609	0.619	0.684

4. EVALUATION ON CHILDREN'S DATA

We evaluated our easiness scores by correlating them against 274 children's performance on next-sentence prediction questions. These questions were inserted randomly by the spring 2003 version of Project LISTEN's Reading Tutor into 179 English-language stories ranging from grades 3-7. None of these stories were in the dataset used to train the coherence measure used to score easiness. The questions asked "Which will come next?" and presented the next three story

[1] https://pytorch.org/docs/stable/optim.htmltorch.optim.lr_scheduler.ReduceLROnPlateau

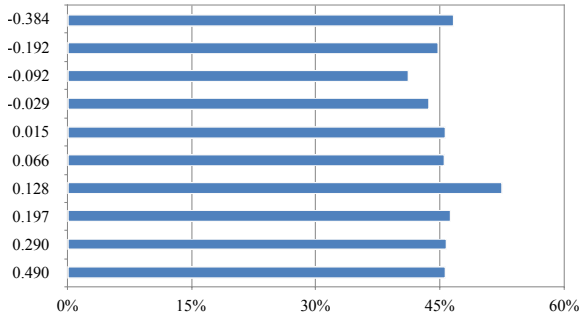

Figure 3: Percentage correct binned by easiness score.

sentences in random order. After data cleaning, we got 1,023 distinct questions with 1,626 responses, of which 45.7% were correct.

344 of these questions had choices with differences in capitalization, as illustrated in Figure 6 (see Appendix). Children might conceivably have used these differences as a clue to eliminate incorrect choices. However, their 622 responses to choices capitalized differently had virtually the same (in fact slightly *lower*) percentage correct (45.5%) as their 1004 responses to choices capitalized the same (45.9%). Evidently children did not make use of this clue. Accordingly, we did not exclude these 622 responses from our dataset.

The questions averaged only 1.59 responses each, far too few to reliably estimate the percentage correct for individual questions. Instead, we split questions by easiness scores into N bins with equal numbers of questions. For N=10, % correct ranged from 41.2% to 52.5%. Figure 3 shows a bar chart with a bar for each of the 10 bins, its average easiness score to its left, and its % correct as its width. The % correct was similar across all 10 bins and unrelated to easiness score. We tried various values of N, ranging from 3 to 128 questions. For each value of N, we correlated the average easiness score of the questions in each bin against their percentage of correct responses. The correlations got weaker as N increased, and were not statistically significant.

To explore why, we regressed response correctness against several features of questions, namely the length and contextual coherence of the correct answer and the two distractors, the length (in characters) of the context, the position of the question in the story (the number of sentences preceding it), and the grade level of the story. We normalized the value of each feature x as $(x - x\_min)/(x\_max - x\_min)$. We performed logistic regression with the normalized feature values for each question as numerical inputs and the correctness of the child's response as binary output. None of the regression coefficients differed significantly from zero. However, their general pattern makes qualitative sense. The contextual coherence of the correct answer was the strongest positive predictor, which makes sense because it measures how well the answer fit the context. The coherence of the harder distractor was the strongest negative predictor, which makes sense because it measures how well that distractor fit the context. The length of the correct answer and the number of preceding sentences in the story were positive predictors, which makes sense because they measure the amount of information provided for selecting the correct answer. Context length and the grade level of the story were negative predictors, which makes sense because reading longer sentences and higher level stories was harder (though better readers read harder stories).

5. CONCLUSIONS

This paper addresses two hypotheses regarding the use of next-sentence prediction questions in assessing children's inter-sentential processing during reading comprehension.

Hypothesis 1: An automated measure of text coherence can predict which of the next 3 sentences will come first. To test hypothesis 1, we trained a BERT-based model of a sentence's coherence with the preceding context to predict whether it comes next. It achieved 61% accuracy on a held-out test set.

Hypothesis 2: An easiness metric based on this measure can predict children's accuracy in selecting the next sentence. To test hypothesis 2, we scored the easiness of the 3-way choice as the coherence of the correct next sentence minus the coherence of the strongest competitor. We then related this score to children's performance on 1,023 such questions presented by Project LISTEN's Reading Tutor to the children while they were using it. There was virtually no correlation. Children answered approximately 45% of the questions correctly regardless of their easiness scores or whether the BERT-based model answered them correctly.

5.1 Limitations and Future Work

If hypothesis 2 were true, we could use a BERT-based coherence model to estimate the difficulty of deciding whether a given sentence will come next in a story context. We could then control question difficulty by using this estimate to help decide which sentence prediction questions to ask. Unfortunately, our results did not support hypothesis 2, which raises the issue of why they did not. The predictor coefficients in our regression analysis to explore this issue made qualitative sense but were not statistically significant.

Perhaps children's performance was affected by the added memory load of considering three sentences as choices. Future work could kid-test the simpler question "Is this next?".

Another possibility is that our coherence model was too impoverished to reflect children's inter-sentential processing. A richer model could capture other aspects such as causal relations, world knowledge, and inference important in story understanding. Or perhaps our BERT model merely needed better adaptation to the domain of children's stories.

An IRT model predicts probability of correctness based on student proficiency minus question difficulty. We did not take direct account of children's differing proficiency, but the Reading Tutor gave children stories at their own reading level, accounting for their proficiency indirectly. Future analyses may need to account for proficiency explicitly.

6. ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments, the children who used Project LISTEN's Reading Tutor, and the team that implemented it and collected our dataset.

7. REFERENCES

[1] J. C. Alderson. Native and nonnative speaker performance on cloze tests. *Language Learning*, 30(1):59–76, 1980.

[2] J. E. Beck, J. Mostow, and J. Bey. Can automated questions scaffold children's reading comprehension? In *International Conference on Intelligent Tutoring Systems*, pages 478–490. Springer, 2004.

[3] N. A. Bond et al. Studies of verbal problem solving: Ii. prediction of performance from sentence-processing scores. technical report no. 87. 1978.

[4] A. Conneau and G. Lample. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[6] Y. Gao, L. Bing, W. Chen, M. Lyu, and I. King. Difficulty controllable generation of reading comprehension questions. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 4968–4974, 2019.

[7] B. S. Hensler and J. Beck. Better student assessing by finding difficulty factors in a fully automated comprehension measure. In *International Conference on Intelligent Tutoring Systems*, pages 21–30. Springer, 2006.

[8] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[9] G. Kurdi, J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204, 2020.

[10] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, 2019.

[11] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pages 807–814. Omnipress, 2010.

[12] H. Nomura. Meaning understanding in machine translation. In *Proc. of Second International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 1988.

[13] D. Porter. The effect of quantity of context on the ability to make linguistic predictions: A flaw in a measure of general proficiency. *Current developments in language testing*, 5(4):63–74, 1983.

[14] T. Shanahan, M. L. Kamil, and A. W. Tobin. Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, pages 229–255, 1982.

[15] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[16] C. Y. Yeung, J. S. Lee, and B. K. Tsou. Difficulty-aware distractor generation for gap-fill items. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 159–164, 2019.

[17] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

APPENDIX

Context: George's favorite subject was math. George learned to be a surveyor of land when he grew up. He joined the army and was a leader during the American Revolution.

Choices	Coherence	Easiness
Correct Answer: He later became the first President of the United States.	0.714	
Distractor 1: George Washington is called the "Father of our Country."	0.470	0.244
Distractor 2: We celebrate his birthday on President's Day in February.	0.310	

Figure 4: A question with easiness score of 0.244.

Context: Both Brad and Sally pointed their flashlights into the dark. All they saw were some spider webs and a dead end. The cave was empty.

Choices	Coherence	Easiness
Correct Answer: Brad felt sad.	0.337	
Distractor 1: He had hoped they would find a big pirate ship or something neat.	0.101	-0.262
Distractor 2: Sally looked around the walls of the cave.	0.599	

Figure 5: A question with easiness score of -0.262.

Context: When all the straw was spun away, and all the bobbins were full of gold. As soon as the sun rose the King came and when he perceived the gold he was astonished and delighted.

Choices	Coherence	Easiness
Correct Answer: But his heart only lusted more than ever after the precious metal.	0.695	
Distractor 1: He had the miller's daughter put into another room full of straw,	0.248	0.447
Distractor 2: much bigger than the first, and bade her, if she valued her life,	0.094	

Figure 6: A question with choices capitalized differently.