Educational Multi-Question Generation for Reading Comprehension

Manav Rathod[1], Tony Tu[2], and Katherine Stasaski[1]
[1]UC Berkeley
[2]Georgia Institute of Technology
[1]{manav.rathod, katie_stasaski}@berkeley.edu
[2]ttu32@gatech.edu

Abstract

Automated question generation has made great advances with the help of large NLP generation models. However, typically only one question is generated for each intended answer. We propose a new task, Multi-Question Generation, aimed at generating multiple semantically similar but lexically diverse questions assessing the same concept. We develop an evaluation framework based on desirable qualities of the resulting questions. Results comparing multiple question generation approaches in the two-question generation condition show a trade-off between question answerability and lexical diversity between the two questions. We also report preliminary results from sampling multiple questions from our model, to explore generating more than two questions. Our task can be used to further explore the educational impact of showing multiple distinct question wordings to students.

1 Introduction

Automatic question generation (QG) is a well-established task in natural language processing. Large generation models have had success producing answer-informed factual comprehension questions, where the intended answer is a span located in a passage (Qi et al., 2020; Wang et al., 2018; Rajpurkar et al., 2016).

Automatically generating factual questions from a passage can benefit students in a reading comprehension environment (Kurdi et al., 2020). However, the majority of question generation work has focused on generating a single question with an intended answer. In a student practice environment, however, it is valuable to have multiple wordings for the same question. This allows students to practice a concept multiple times without encountering identical language. This additionally allows a wording of the question to be held out for assessment.

We propose a new question generation task, *multi-question generation*, which takes as input an intended answer and produces both (1) an initial question and (2) n reworded questions which maintain the semantic meaning of the original question while varying language used. Although multiple questions can be generated about different concepts pertaining to an intended answer, we specifically aim to generate questions which assess knowledge of the same concept, varying only the language used in the question.

Another issue with current generation systems is the large overlap in words between the reading passage and generated question. Table 1 shows an example of an undesired output from a current question generation system. Note the large overlap between the generated question and input passage, allowing students to scan for the answer in the paragraph. For our task, we additionally specify that the text of resulting questions should differ from the content passage.

We propose automatic metrics grounded in desirable properties of the generated set of questions. Because we intend the questions to have the same intended answer, we measure both (1) whether a *question answering* (QA) model is able to produce the correct answer for each question (Yuan et al., 2017) and (2) the semantic similarity between the generated questions, measured using SBERT (Reimers and Gurevych, 2019). Also, because we intend for the questions to have distinct wordings, we propose using a known n-gram overlap metric, Paraphrase In N-Gram Changes (PINC), between pairs of questions (Chen and Dolan, 2011). Finally, because each generated question is tied to a passage, we propose using PINC to compare overlap between each question and the input passage.

We report results using a variety of question generation conditions, including a paraphrase model, a QG model fine-tuned to generate two questions, and the use of decoding constraints to improve question wording diversity. Our publicly-released code and generated questions can be used to ex-

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 216 - 223

Original Passage	The Sarah Jane Adventures, starring Elisabeth Sladen **who reprised her role as investigative journalist Sarah Jane Smith**, was developed by CBBC; a special aired on New Year's Day 2007 and a full series began on 24 September 2007. A second series followed in 2008, notable for (as noted above) featuring the return of Brigadier Lethbridge-Stewart. A third in 2009 featured a crossover appearance from the main show by David Tennant as the Tenth Doctor. In 2010, a further such appearance featured Matt Smith as the Eleventh Doctor alongside former companion actress Katy Manning reprising her role as Jo Grant. A final, three-story fifth series was transmitted in autumn 2011 uncompleted due to the death of Elisabeth Sladen in early 2011.
Answer	Elisabeth Sladen
Question	**who reprised her role as investigative journalist sarah jane smith**?

Table 1: Example passage taken from SQuAD dataset with corresponding question generated from ProphetNet (Qi et al., 2020). Bolded text shows overlap between the input passage and generated question, which is not desired.

plore the impact of integrating these questions into educational applications[1].

2 Related Work

2.1 Question Generation

Many question generation models are fine-tuned from large language models, achieving considerable success in producing factual reading comprehension questions (Grover et al., 2021; Chan and Fan, 2019; Sultan et al., 2020). Other work aims to generate questions given passages from educational textbooks (Wang et al., 2018; Stasaski et al., 2021). However, these QG models are trained to only produce a single question from a context paragraph and intended answer.

2.2 Educational Question Application

Anderson and Biddle (1975) find that asking factual questions during reading can aid in the ability to recall a story. Furthermore, providing students with multiple phrasings of the same question has the potential to ensure students have fully mastered a concept (Kurdi et al., 2020). Rephrasing a question when students answer incorrectly has been included in best practices for educational question asking (Tofade et al., 2013) as well as a component of Elaborative Feedback (Murphy, 2007). Additionally, past educational research has also found that providing a human-written paraphrased wording of the same question has been shown to improve reading comprehension of students who are less skilled compared to a baseline with only one question wording (Cerdán et al., 2019).

Following this past educational work, we propose leveraging neural systems to generate multiple diverse question wordings. Our new task allows future work to study this at scale.

3 Multi-Question Generation

Given the potential educational benefits that come from answering questions with different wordings, we propose *Multi-Question Generation*, with the goal of producing multiple semantically similar, lexically diverse questions with the same intended answer. An intended answer and a passage are the input to the task while the multiple diversely-worded questions are the output. Although an intended answer can have multiple concepts with which questions can be generated from, these multiple questions should assess the same concept. An example of this can be seen in Table 2.

3.1 Evaluation

We propose an evaluation framework to assess the quality of the generated questions. Because we do not have a gold human-collected dataset of rephrased questions, we propose heuristic evaluation metrics. We evaluate the generated questions using a combination of PINC, a QA model, and SBERT cosine similarity.

Because the set of questions should have limited lexical overlap, we use PINC to measure the n-gram overlap among pairs of questions (Chen and Dolan, 2011). Specifically, for two generated questions q_1 and q_2, the PINC score is calculated

[1]Code is available at https://github.com/kstats/MultiQuestionGeneration.

Original Passage	Victoria (abbreviated as Vic) is a state in the south-east of Australia. Victoria is Australia's most densely populated state and its second-most populous state overall. Most of its population is concentrated in the area surrounding Port Phillip Bay, which includes the metropolitan area of its capital and largest city, Melbourne, which is Australia's second-largest city. Geographically the smallest state on the Australian mainland, Victoria is bordered by Bass Strait and Tasmania to the south,[note 1] New South Wales to the north, the Tasman Sea to the east, and South Australia to the west.
Answer	second-largest
Question 1	where does melbourne rank in terms of the size of cities in australia?
Question 2	what is melbourne's population status?

Table 2: Selected *Two-Question Generation* output from the 2QG No Question Trigram model, presented in Section 4.

as:

$$PINC(q_1, q_2) = \frac{1}{N}\sum_{n=1}^{N} 1 - \frac{|\text{n-gram}_{q_1} \cap \text{n-gram}_{q_2}|}{|\text{n-gram}_{q_2}|}$$

where N is the maximum n-gram considered and n-gram$_{q_1}$ and n-gram$_{q_2}$ are the list of n-grams in the first and second questions, respectively.

However, since this metric is not symmetric and there is no reason to treat one question as the standard over another, we compute the score in both directions and average:

$$PINC_{sym}(q_1, q_2) = \frac{PINC(q_1, q_2) + PINC(q_2, q_1)}{2}$$

We use $PINC_{sym}$ to calculate distinction among the set of generated questions Q for a given example as: $\forall_{q_i,q_j \in Q\,:\,i \neq j} PINC_{sym}(q_i, q_j)$.

We additionally propose using PINC to calculate the distance from each question to the context paragraph C: $\forall_{q_i \in Q} PINC(C, q_i)$. Note that here we use the asymmetric PINC since we want to explicitly reward the question for introducing new n-grams not found in the context paragraph.

We calculate PINC up to trigrams, manually confirming this to balance allowing important phrases to be restated when appropriate without allowing for long copied phrases.

Next, we draw from past work which has used Question Answering models to evaluate the accuracy of Question Generation systems (Yuan et al., 2017). Following this, we use the performance of a Question Answering model[2] to ensure the generated questions are answerable. For measuring QA accuracy, we use a macro-averaged F1, treating the predicted answer and ground truth as bags of tokens, as done in the original SQuAD paper (Rajpurkar et al., 2016).

Lastly, we aim to measure the semantic similarity between generated questions to ensure that the questions assess the same content. To do this, we use a pre-trained SBERT model[3] (Reimers and Gurevych, 2019) to encode each question into an embedding and take the cosine similarity between each pair of embeddings.

4 Experimental Conditions

We begin with the task of generating two questions (results of generating more than two questions can be seen in Section 6). To approach this task, we leverage a high-quality neural question generation model, ProphetNet (Qi et al., 2020). In order to generate multiple questions, we explore (1) transforming ProphetNet's single question output into a paraphrased second question, (2) fine-tuning ProphetNet to output two questions sequentially, and (3) sampling multiple times from ProphetNet.

All models use beam search with a beam size of 10 unless otherwise stated. For sampled results, we use nucleus sampling (Holtzman et al., 2020) with $p = 0.95$. All results are reported on the SQuAD 1.1 development set (Rajpurkar et al., 2016).

[2] https://huggingface.co/bert-large-uncased-whole-word-masking-finetuned-squad

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

4.1 Question Paraphrasing

For our first approach (1QG+Para), we use the ProphetNet model to generate a single question given the context and answer. We then pass this generated question into a paraphrasing model. Ideally, paraphrasing the original input will preserve the meaning of the question while modifying the lexical content. We use a T5-based model (Raffel et al., 2020) that is trained on the Quora Question (Chen et al., 2018) pairs dataset[4].

4.2 Two-Question Generation

The previous approach (1QG+Para) is not ideal because the paraphrase model does not have access to the context or intended answer. Thus, we finetune ProphetNet to output two question for any given context and answer (2QG). We finetune ProphetNet on a dataset of paraphrased questions created from 1QG+Para's paraphrase model. We augment each training example in the SQuAD training dataset with an additional paraphrase and then finetune ProphetNet to predict a sequence of two questions separated by a separator token, "[X_SEP]."

We fine-tuned ProphetNet for 10 epochs using a learning rate of 1e-5 using the Adam Optimizer (Kingma and Ba, 2015) on the entire SQuAD training set. We initialized the model using the weights from Transformers (Wolf et al., 2020) [5]. We trained on an NVIDIA Titan RTX for 2 days.

4.2.1 Constrained Generation

While 2QG is able to output two well-formed questions, its ability to vary lexical diversity may be limited by the training data. To further encourage the model to output different questions, we add constraints to the 2QG model's generation process to force this property. We explore two constraints: 1) requiring the generated questions to not repeat any trigrams across both questions (2QG No Question Trigram) and 2) requiring the generated questions to not repeat any trigrams from the input passage (2QG No Context Trigram). We also explore a version of ProphetNet which has both of these constraints (2QG No Question-Context Trigram).

4.3 Sampling

Finally, we explore potential questions which can be uncovered by sampling from the QG model's learned distribution. For the 1QG case, we sample from ProphetNet twice to generate the two questions (1QG 2-Sample). For the 2QG model (2QG Sample), we sample once as the model output already contains two well-formed questions. In Section 6, we explore sampling from the 2QG model more than once.

[4] https://huggingface.co/ramsrigouthamg/t5_paraphraser

[5] https://huggingface.co/docs/transformers/model_doc/prophetnet

5 Two-Question Generation Results

Two-Question Generation Results can be found in Table 3. Appendix A contains randomly-sampled model output for one of the best-performing models, 2QG No Question Trigram.

We observe that restricting the repetition of trigrams in the question generation increases the PINC score, which is expected as generating repeating trigrams is constrained. However, this comes at the cost of having a lower QA score.

We also note higher QA scores for the first question compared to the second, meaning answerability might be less important when rephrasing the initial question. The drop in performance from QA1 to QA2 for 1QG+Para is anticipated as the paraphrase model does not have access to the answer or context passage. However, surprisingly, we observe similar performance drops with 2QG models (in particular 2QG No Question Trigram). The gap in quality is increased when the PINC score between the questions is higher, indicating a tradeoff between PINC score and QA score. We also observe that restricting the trigrams from the *context paragraph* (2QG No Context Trigram) increases the PINC score with respect to the context paragraph as expected, but does so at a smaller cost to the QA score.

Lastly, we note an inverse relationship between inter-question PINC score and SBERT similarity. This indicates that diversifying lexical content of questions may come at the cost of maintaining semantic similarity between the two questions.

6 Toward Multi-Question Generation

We next explore bridging the gap from Two-Question Generation to Multi-Question Generation. While the 2QG model was fine-tuned to produce two questions sequentially, we explore the extent to which sampling from this model can produce sets of more questions. We take the 2QG model and sample from it multiple times using nucleus sampling (p=0.95). We consider sets of 2, 4, 6, and 8 questions.

Model	Q1-Q2	C-Q1	C-Q2	QA1	QA2	SBERT
1QG 2-Sample	0.32	0.49	0.50	**0.83**	**0.82**	0.91
1QG+Para	0.33	0.49	0.58	**0.83**	0.63	**0.98**
2QG	0.11	0.57	0.59	0.82	0.80	**0.98**
2QG Sample	0.12	0.58	0.60	**0.83**	0.80	**0.98**
2QG No Context Trigram	0.12	**0.75**	0.75	0.79	0.77	**0.98**
2QG No Question Trigram	**0.77**	0.58	0.76	**0.83**	0.63	0.83
2QG No Question-Context Trigram	**0.77**	**0.75**	**0.80**	0.79	0.62	0.85

Table 3: Two-Question Generation results. Models explored are discussed in Section 4. The first three columns report the PINC score between the first question (Q1), the second question (Q2), and the context (C). The next two columns report the QA model's F1 score for the first (QA1) and second (QA2) generated question. The last column reports the SBERT cosine similarity between the generated questions. Higher values are better for all metrics.

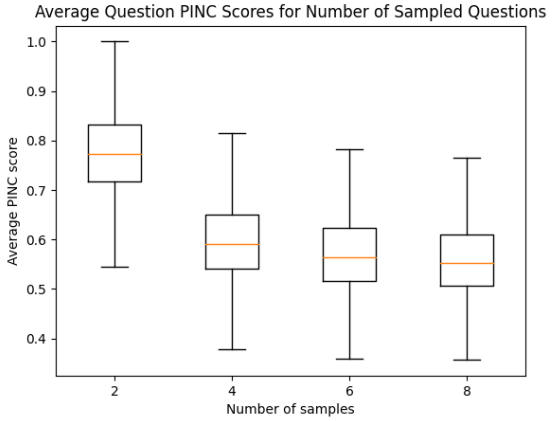

Figure 1: Average PINC between-question scores for increasing number of question samples.

We examine the average between-question PINC scores for the generated question sets, to explore whether sampling can uncover unique question wordings. Results can be seen in Figure 1. We find a sharp decline in PINC score for more than two questions. Future work should explore other ways of generating more than two questions.

7 Future Work and Conclusion

Although automated evaluation metrics can measure the desirable properties of our Two-Question Generation model outputs at scale, they are also limited. Future work could include human evaluation metrics to measure the semantic quality and lexical diversity more robustly.

Future work should also explore using desirable question metrics in a reinforcement learning objective to produce higher quality questions, similar to previous work in abstractive summarization (Laban et al., 2020) and text simplification (Laban et al., 2021).

Additionally, more advanced paraphrase systems, such as the syntax-aware system proposed in Kumar et al. (2020), could be leveraged for our task. This work can explore which syntactic exemplars can be leveraged to generate questions with varying syntactic structure.

Additionally, future work should also include teacher evaluation to collect education-specific feedback on sets of questions and our desirable question properties. This work can help better define what constitutes a good question and potentially uncover different automated metrics.

Future work can leverage our task to evaluate the educational impact of multiple diverse question wordings. Multi-Question Generation can be integrated into a reading comprehension environment to test student reactions to a reworded question. Generating multiple question wordings can fully test the students' reading comprehension and ability to apply information in new situations. Our publicly-released pipeline has the potential to generate multiple wordings of the same questions to enrich educational resources at scale.

Acknowledgements

This work was supported by an AWS Machine Learning Research Award and an NVIDIA Corporation GPU grant. We thank the three anonymous reviewers as well as Marti Hearst, Emily Xiao, Philippe Laban, and the Hearst Lab research group for their useful comments.

References

Richard Anderson and Barry Biddle. 1975. On asking people questions about what they are reading. *Psychology of Learning and Motivation*, 9.

R. Cerdán, A. Pérez, E. Vidal-Abarca, and J. F. Rouet. 2019. To answer questions from text, one has to understand what the question is asking: differential effects of question aids as a function of comprehension skill. *Reading and Writing*, 32(8):2111–2124.

Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 154–162, Hong Kong, China. Association for Computational Linguistics.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA. Association for Computational Linguistics.

Zihan Chen, Hongbo Zhang, Xiaoji Zhang, and Leqi Zhao. 2018. Quora question pairs. *URL https://www. kaggle. com/c/quora-question-pairs*.

Khushnuma Grover, Katinder Kaur, Kartikey Tiwari, Rupali, and Parteek Kumar. 2021. Deep learning based question generation using t5 transformer. In *Advanced Computing*, pages 243–255, Singapore. Springer Singapore.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. 2020. A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education*, 30(1):121–204.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6365–6378, Online. Association for Computational Linguistics.

Philip Murphy. 2007. Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning & Technology*, 11(3):107–129.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Katherine Stasaski, Manav Rathod, Tony Tu, Yunfang Xiao, and Marti A. Hearst. 2021. Automatically generating cause-and-effect questions from passages. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 158–170, Online. Association for Computational Linguistics.

Md Arafat Sultan, Shubham Chandel, Ramón Fernandez Astudillo, and Vittorio Castelli. 2020. On the importance of diversity in question generation for QA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5651–5656, Online. Association for Computational Linguistics.

Toyin Tofade, Jamie Elsner, and Stuart T. Haines. 2013. Best practice strategies for effective use of questions as a teaching tool. *American journal of pharmaceutical education*, 77(7):155–155. 24052658[pmid].

Zichao Wang, Andrew S. Lan, Weili Nie, Andrew E. Waters, Phillip J. Grimaldi, and Richard G. Baraniuk. 2018. Qg-net: A data-driven question generation model for educational content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, L@S '18, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xingdi Yuan, Tong Wang, Caglar Gulcehre, Alessandro Sordoni, Philip Bachman, Saizheng Zhang, Sandeep Subramanian, and Adam Trischler. 2017. Machine comprehension by text-to-text neural question generation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 15–25, Vancouver, Canada. Association for Computational Linguistics.

A Generated Output

Context	Answer	Q1	Q2	Q1-Q2	C-Q1	C-Q2
The annual NFL Experience was held at the Moscone Center in San Francisco. In addition, "Super Bowl City" opened on January 30 at Justin Herman Plaza on The Embarcadero, featuring games and activities that will highlight the Bay Area's technology, culinary creations, and cultural diversity. More than 1 million people are expected to attend the festivities in San Francisco during Super Bowl Week. San Francisco mayor Ed Lee said of the highly visible homeless presence in this area "they are going to have to leave". San Francisco city supervisor Jane Kim unsuccessfully lobbied for the NFL to reimburse San Francisco for city services in the amount of $5 million.	$5 million	how much did kim ask the nfl to reimburse san francisco for city services during the super bowl?	what did lee ask for from the nfl in terms of financial assistance for san francisco during the superbowl?	0.79	0.42	0.75
Newcastle has three cathedrals, the Anglican St. Nicholas, with its elegant lantern tower of 1474, the Roman Catholic St. Mary's designed by Augustus Welby Pugin and the Coptic Cathedral located in Fenham. All three cathedrals began their lives as parish churches. St Mary's became a cathedral in 1850 and St Nicholas' in 1882. Another prominent church in the city centre is the Church of St Thomas the Martyr which is the only parish church in the Church of England without a parish and which is not a peculiar.	Coptic	what is the third cathedral in newcastle?	what are the three cathedrals of newcastle?	0.86	0.65	0.71
With Rivera having been a linebacker with the Chicago Bears in Super Bowl XX, and Kubiak replacing Elway at the end of the Broncos' defeats in Super Bowls XXI and XXIV, this will be the first Super Bowl in which both head coaches played in the game themselves.	Super Bowl XX	what was the first super bowl in which both head coaches played?	in what way did the first superbowl ever take place?	0.84	0.18	0.86
Cultural imperialism is when a country's influence is felt in social and cultural circles, i.e. its soft power, such that it changes the moral, cultural and societal worldview of another. This is more than just "foreign" music, television or film becoming popular with young people, but that popular culture changing their own expectations of life and their desire for their own country to become more like the foreign country depicted. For example, depictions of opulent American lifestyles in the soap opera Dallas during the Cold War changed the expectations of Romanians; a more recent example is the influence of smuggled South Korean drama series in North Korea. The importance of soft power is not lost on authoritarian regimes, fighting such influence with bans on foreign popular culture, control of the internet and unauthorised satellite dishes etc. Nor is such a usage of culture recent, as part of Roman imperialism local elites would be exposed to the benefits and luxuries of Roman culture and lifestyle, with the aim that they would then become willing participants.	Roman	what culture is an example of cultural imperialism?	what is cultural imperialism and what are some examples of this?	0.77	0.70	0.77
BSkyB's standard definition broadcasts are in DVB-compliant MPEG-2, with the Sky Movies and Sky Box Office channels including optional Dolby Digital soundtracks for recent films, although these are only accessible with a Sky+ box. Sky+ HD material is broadcast using MPEG-4 and most of the HD material uses the DVB-S2 standard. Interactive services and 7-day EPG use the proprietary OpenTV system, with set-top boxes including modems for a return path. Sky News, amongst other channels, provides a pseudo-video on demand interactive service by broadcasting looping video streams.	Dolby Digital	what kind of soundtracks are optional on sky movies and sky box office?	what kinds of soundtracks do sky sky box offices and sky movies use?	0.69	0.46	0.70

Table 4: Randomly-sampled model output from the 2QG No Question Trigram model.