Distraction Generation using Google T5: A Study

Fernando Gonzalez Adauto
ETH Zürich
fgonzalez@student.ethz.ch

Kevin Golan
ETH Zürich
kgolan@student.ethz.ch

Abstract

Distraction Generation (DG) is the remaining task in Automated Quiz Generation that is yet to be mastered by the state-of-the-art in NLP. Given a reference text and question, an ideal DG model produces a set of distractors that are contextually coherent but factually wrong. The literature presents a plethora of approaches that address this problem. But none, so far, provide a robust framework that consistently produces distractors of quality. Our contribution to this field is twofold: (1) We are the first to apply Google's T5 model specifically to this problem. (2) We modify T5's loss function by introducing a cosine similarity term to generate factually wrong distractors. Running T5-small and T5-base, we obtain results that fall short of established baselines but are nonetheless encouraging. Additionally, our results suggest that the modified loss function improves performance.

1 Introduction

The multiple-choice question (MCQ) test is a long-established assessment format dating back to the early 20th century. At first, it was used as a tool to assess a teacher's knowledge on a topic, whose performance would determine his or her salary (U. of Texas, 2001). Since then, MCQ's, as a form of assessment, have become ubiquitous in almost all levels of education.

A typical MCQ test will likely follow one of two formats. The first type consists of directly assessing the candidate's knowledge on a topic by posing several questions relating to that subject, each accompanied with multiple candidate answers. Among the proposed candidates, typically, only one is correct. The second MCQ test format, which will be of the type considered in this paper, consists of providing the candidate with a reference context – be it a chapter, a paragraph, a table – upon which multiple-choice questions are formulated. Once again, in almost all instances, only one of the proposed candidates is the correct one.

As learning is increasingly shifting online and the proliferation of online academic resources is far from slowing down, new avenues to measure feedback of one's digestion of information are being sought. This trend can be credited for the emergence of alternative learning formats such as Massive Open Online Courses (MOOCs), tutorial blogs, coding platforms, and more. In this paper, we aim to contribute to the field of MCQ Quiz Generation by focusing on the problem of Distraction-Generation (DG).

Distraction Generation deals with the task of generating candidate choices that are *incorrect* but *plausible*, given a question and context. Although sophisticated methods to produce question-answer pairs from a given text already exist, using, for example, BERT or custom-made Transformer architectures - distraction generation remains an underdeveloped area of research.

Common challenges found in Distraction Generation include, but are not limited to: distractors that completely ignore a question's context; distractors that make little grammatical sense; distractors that express the answer in alternative ways (hence are not incorrect), and distractors that can be disregarded through simple reasoning. Although several interesting approaches have been proposed to mitigate some of these issues (discussed in the ensuing Literature Review section), we are yet to come across a robust, generalized blueprint for DG that addresses most if not all of them.

With the propositions and results put forward by this paper, we hope to provide a robust framework that produces distractors of quality and addresses the aforementioned challenges in DG. We intend to tackle those using Google's T5 architecture and by implementing our custom loss function. By undertaking such research, we hope to contribute to the ambitious goal of making high-quality quiz generation a possibility at scale.

2 Related Work

To the best of our knowledge, research in DG has seen two streams so far: short-form DG and long-form DG.

Short-form DG typically appears in close-style MCQ generation, which produces questions whose answers are single words as found in Figure 1. Although this method can yield effective distractors (using relatedness between words or external knowledge banks, for example), it ultimately restricts the type of questions that could be produced, thus restricting the depth of knowledge assessed. Due to the shortcomings of short-form DG, we turn instead to long-form DG, which will be the centre of our work and of which an instance can be found in Figure 2. Here, MCQ's are generated by pairing questions with options of variable length. Relevant findings in both approaches are discussed in the ensuing paragraphs.

2.1 Short-Form Distraction Generation

Multiple approaches have been proposed to solve the task of short-form DG. Among others, applying variations on the grammar and vocabulary of the ground truth (Hoshino and Nakagawa, 2007); using orthographically, morphologically, or phonetically similar words (Pino and Eskénazi, 2009); selecting distractors based semantic similarity using Word2Vec (Jiang and Lee, 2017), and using external knowledge banks to propose tokens that are thematically related to the ground truth (Ren and Zhu, 2020). Despite the merit of these approaches, none of these approaches tests a candidate's reading comprehension ability but instead focus more on grammatical or factual knowledge.

Question: What is the capital of Spain?

Possible Answers:
a) Barcelona
b) Paris
c) Brussels
d) Madrid

Figure 1: Example of a question containing short-form distractors.

Context:
[…] In Panama, the Kuna people saved their forest. They made a forest park which tourists pay to visit. The Gavioes people of Brazil use the forest but protect it as well. […]

Question:
Those people built roads and airports in order to _.

a) carry away the gold conveniently
b) make people there live a better life
c) stop spreading the new diseases
d) develop the tourism there.

Figure 2: Example of a question containing long-form distractors. Selected from RACE and slightly modified for clarity purposes.

2.2 Long-Form Distraction Generation

When surveying the research in long-form DG, we found that – at a higher level at least – most groups incorporated some form of Transformer architecture in dealing with DG, and a minority used Bi-LSTM or RNN-based architectures. Within the group that employs Transformer architectures, we noticed a split in the practice of long-form DG, best explained by two streams. The first stream, which encompasses most of long-form DG research, proceeds in the task by fine-tuning derived versions of the model proposed by Vaswani et al. (2017) such as BERT, T5, and PEGASUS (Zhang et al., 2020; Devlin et al., 2019; Raffel et al., 2020a). The second stream, which you find less often, encompasses methods that opted to use a custom-made architecture for this task but still uses a variant of the Transformer architecture.

2.3 Custom Transformer Architectures

Attention mechanisms have been a keystone in Natural Language Processing (NLP). Many works use this mechanism as the principal component of their architecture to build specifically designed models to solve DG. One of the first works in DG proposed a hierarchical encoder-decoder framework (HSA) with static and dynamic attention mechanisms to address both word and sentence importance (Gao et al., 2018). Then Zhou et al. (2019) propose a Co-attention Hierarchical Network in which they use article-to-question and question-to-article attention layers to allow the encoder to capture interactions between article and question. Maurya and Desarkar (2020) use HMD-Net, which consists of one encoder and three decoders with a dissimilarity loss. Finally, Qiu et al. (2020) propose EDGE, in which they introduce two modules that use attention to reform the passage and the question. They generate multiple distractors using beam search and control the distance among them.

2.4 Variations of Encoder-Decoder Architectures

When looking at the literature that uses derivations of the transformer suggested by Vaswani et al. (2017), we found long-form DG strategies involving BERT and its descendants; Google's T5 and PEGASUS; and combinations of GPT2 and BERT.

Offerijns et al. (2020) propose an interesting strategy in which they split the task of distraction generation into two parts: distraction generation and then distractor filtering. They generate distractors using GPT2 (Radford et al., 2019) and thereafter filter them using DistilBERT(Sanh et al., 2020). By applying this filtering mechanism, they approach the problem of distractor quality, i.e. assessing if the distractor fits the question's context and whether it can be considered a grammatically sound answer. Additionally, they apply a repetition penalty that "punishes" their model for generating similar texts, enforcing the generation of syntactically dissimilar distractors.

Finally, Lelkes et al. (2021) tackle Distraction Generation using T5 and PEGASUS models. Using a pre-trained model, they generate distractors by taking the output when prompted to answer a question. Using this for direct question-answering, they observed that they could obtain an arbitrary amount of sample answers if they fed an arbitrary number of questions. After obtaining experimental results, they conclude with a case study that found that for a significant majority of questions, at least one distractor generated by T5 was considered plausible to a human reader.

3 Background

3.1 Task definition

The goal of the DG task can be formulated as follows: given a text passage P, a question Q generated from the passage and an answer A to that question, generate the best wrong option D. Formally, we aim to find the best distractor D that maximizes the conditional log-likelihood given P, Q, and A:

$$\tilde{D} = \arg\max_{D} \log P(D \mid P, Q, A) \qquad (1)$$

T5 Configurations Breakdown

Model Version	Parameter Count (millions)
T5 Small	60
T5 Base	220
T5 Large	770
T5-3B	3000
T5-11B	11000

Table 1: Google T5 configurations and associated parameter count.

3.2 T5 Model Overview

We use Google's T5 model to generate distractors from an input text sequence and question, each represented as tokens. As previously discussed in our Related Work section, T5 is a slightly modified variant of the architecture proposed by Vaswani et al. (2017). Additionally, the T5 repository offers multiple model configurations (or versions) of its model with varying numbers of pre-trained parameters, details of which are in Table 1.

Google T5's structure relies on two main components: an encoder and a decoder. As described in Raffel et al. (2020a), the workflow in T5 runs as follows: a tokenized text sequence is inputted into the architecture and then converted to an embedding representation. This embedding representation then passes into the architecture's encoder that then transfers its output to the decoder. Finally, the decoder transfers its output sequence to a fully connected layer that produces a softmax output. Figure 31, hereunder, is a flowchart describing how we apply T5 to our work.

3.2.1 Encoder

The encoder is organized as a stack of modules (12 to be precise), each composed of two components: a self-attention layer and a fully connected layer. Once an output representation exits a module, it is combined with a residual skip connection that adds the module's input to its output. After passing through the encoder's blocks, the resulting sequence is forwarded into the decoder.

3.2.2 Decoder

The decoder operates in a similar fashion, containing 12 modules as well, but with an important architectural tweak: an additional attention mechanism is introduced between the self-attention layer and the fully-connected layer.

The function of the attention mechanism in T5's decoder distinguishes itself in two ways. First, it operates in a auto-regressive manner, meaning that it only attends to *past* outputs. And second, each attention mechanism is split into different ”heads”, whose outputs are then concatenated before further processing.

3.2.3 Loss Function & Optimization

T5 was conceived for text-to-text tasks. Its training is grounded on teacher forcing, i.e. using the ground-truth from prior steps as inputs to present or future iterations. To that end, T5 runs the standard maximum likelihood and the cross-entropy loss to train its network's weights (Raffel et al., 2020a).

The maximum likelihood's definition and adaptation to our problem is found in (1). The cross-entropy loss is defined hereunder (Raffel et al., 2020b): given a prediction *x* and a class *C* belonging to some vocabulary set *V*, the cross-entropy loss is defined as:

$$\mathrm{L}(x, C) = -\ln\left(\frac{\exp(x[C])}{\sum_j \exp(x[j])}\right) \forall j \in V \quad (2)$$

$$\mathrm{L}(x, C) = -x[C] + \ln\left(\sum_j \exp(x[j])\right) \quad (3)$$

During each training epoch, the cross-entropy loss is averaged over its minibatch size. Then, weights are updated using the AdaFactor optimizer (Shazeer and Stern, 2018).

4 Method

4.1 Model Prefix Definition

Our approach similar to Lelkes et al. (2021) consists of fine-tuning T5. However, in our work, we fine-tune T5 with a new prefix “distraction” and concatenate to the input: the passage, the question, and the answer in the following format: “dist *Q*: question; *A*: answer; *P*: passage”. Each one of the distractors of each question is used as a target sequence, thus generating three input-output pairs per question. Then, we use the pre-trained tokenizer corresponding to the T5 pre-trained model to get the numerical input representations.

4.2 T5 Version Choice & Resource Constraints

As a baseline, we ran T5-small and T5-base with the new ”distraction” prefix. Because of resource limitations, we were unable to run T5-base in the Leonhard cluster. We thus decided to run it once using Google Colab, and then performed the rest of the experiments using T5-small.

4.3 Modifications to the Loss Function

One of the main problems found in past approaches in DG is high similarity between the generated distractors and the answer. To address that, we performed experiments using a two-term loss function. The first component is the standard cross-entropy loss. The second term is a cosine similarity loss that penalizes the distractors' (D) similarity with the correct answer (A).

$$\cos(A, D) = \frac{A \cdot D}{\|A\| \cdot \|D\|} \quad (4)$$

4.3.1 Computing the Loss Function

To compute the first term, we first take the output logits from T5's decoder and obtain the predicted text sequence by applying the softmax function. We then pass that output through the encoder and use the final hidden state as the distractor representation. Similarly, we forward the ground truth (the answer) through the encoder and take the last hidden state as its representation. With those two representations, we then compute the cosine similarity and multiply the result by some weight λ to then add it to the cross-entropy loss using a convex combination of the two terms. The diagram in Figure 3 shows how we compute the loss function. The parameter λ is a tunable hyper-parameter for our model. Hence, our optimization loss function presents itself as follows.

$$L = (1 - \lambda)\mathrm{L}(x, D) + \lambda \cos(A, D) \quad (5)$$

5 Evaluation

5.1 Datasets

We are using the reading comprehension dataset RACE (Lai et al., 2017). Each question is paired with a single paragraph passage, the correct answer, and three distractors. The dataset was collected from English examinations in China, which were designed for middle school and high school students. The dataset is already split in training, validation and test set. Moreover, we use the preprocessed dataset proposed by Gao et al. (2018). There, they remove a large quantity of fill-in-the-blank questions and distractors that have no semantic relevance with the article (the statistics of the

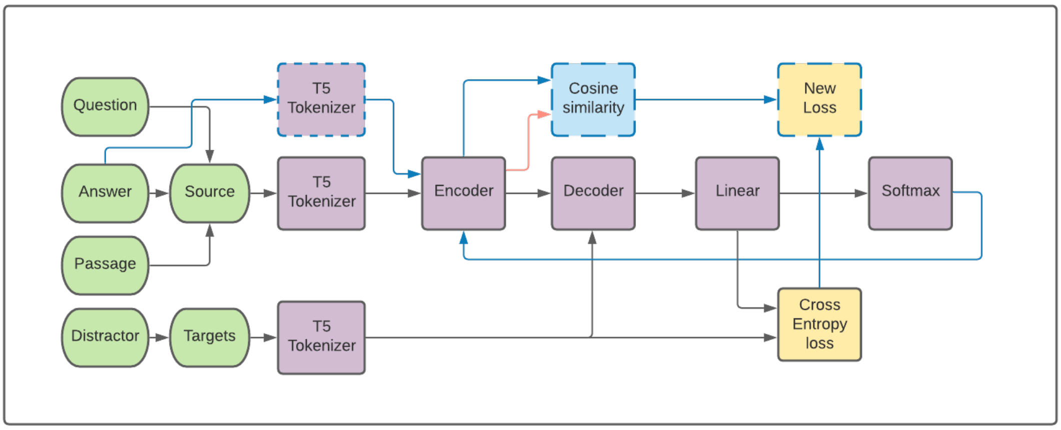


Figure 3: An overview of the model and the process to compute the new loss function. The blocks required to compute the extra term in the loss function are represented with dotted lines

	BLEU 1	BLEU 2	BLEU 3	BLEU 4
T5-small	24.34	8.03	4.50	3.44
T5-small + new loss	25.36	9.09	5.18	3.86
T5-base	29.14	11.71	6.97	5.06
HSA	27.32	14.69	9.29	6.47
Co-Att	28.65	15.15	9.77	7.01
HMD-Net	30.99	17.30	11.09	7.52
EDGE	33.03	18.12	11.35	7.57
BDG	39.81	24.81	17.66	13.56

Table 2: Performance comparison results for 1st distractor our models vs HSA (Gao et al., 2018), Co-Att(Zhou et al., 2019), HMD-Net (Maurya and Desarkar, 2020), EDGE (Qiu et al., 2020), BDG (Chung et al., 2020)

	Validation	Train	Test
# Questions	5681	41505	5779
# Distractors	16132	129226	16266

Table 3: Processed dataset statistics

dataset are shown in Table 3). It is important to note, however, that after pre-preprocessing, the resulting dataset's size is significantly smaller than the original (around half the number of distractors are removed). Nonetheless, having experimented with both, we find that training (given the limited number of resources at hand), on the preprocessed dataset results in better performance.

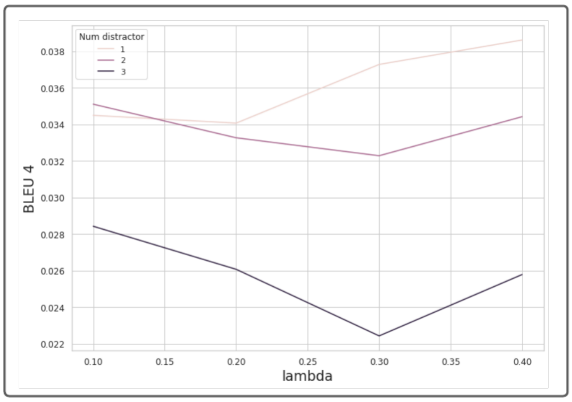


Figure 4: BLEU 4 score for different values of λ

5.2 Results

5.2.1 T5 Performance Overall

We evaluated our models (T5-small and T5-base with new prefix and, T5 new loss) using BLEU (1 to 4) to compare each one of our 3 predicted distractors with the 3 ground truth distractors. The distractors are generated via beam search using the utilities provided in the Huggingface framework (Wolf et al., 2020). In table 4 and 5 we show the evaluation of our 2 best models and in table 2 the comparison with previous approaches. We ran one iteration of T5-base because of resource constraints and this model achieved the best per-

T5 Small (New Loss)				
	BLEU 1	**BLEU 2**	**BLEU 3**	**BLEU 4**
Distractor 1	25.36	9.09	5.18	3.86
Distractor 2	24.69	8.28	4.66	3.44
Distractor 3	19.52	6.34	3.34	2.57

Table 4: Average BLEU Score for 3 distractors produced by T5 Small with modified loss function.

T5 Base				
	BLEU 1	**BLEU 2**	**BLEU 3**	**BLEU 4**
Distractor 1	29.14	11.71	6.97	5.06
Distractor 2	30.18	11.28	6.37	4.66
Distractor 3	26.70	9.47	5.20	3.82

Table 5: Average BLEU Score for 3 distractors produced by T5 Base with original loss function.

formance among our models. Otherwise, our results fell short of previous approaches that we've established as baselines. Nonetheless, our results suggest that with more resources available, our performance could be further improved using larger batch sizes, more epochs and larger models like T5-large, T5-3b or T5-11b.

5.2.2 Modification to the Loss

In evaluating T5 with the modified loss function, we trained the model with different values of the parameter λ. We evaluated the performance of our model using the validation set whose results, for different lambda, are found in Figure 4. We observe that the BLEU-4 score improves with increasing values of λ, and that it edges T5-small's performance when using just the cross entropy loss function. This suggests that the inclusion of this penalty term improves the performance of our model.

In Table 6 and Figure 8 we show the complete performance table with scores for the 3 distractors and for different values of λ. We chose BLEU-4 as the main metric because it reflects better the similarity between 2 long texts since it takes into account groups of 4 words. We selected the model with λ =0.4 as the best one because it shows better performance for the 1st distractor in BLEU 2,3 and 4. In Figure 8 we observe that the score for the 2nd and 3rd distractors doesn't improve for larger values of λ as it does with the 1st distractor. One of the reasons why the cosine term is not very helpful for 2nd and 3rd distractors could be the fact that the model is trained to generate the best distractor and this training process doesn't take into account the way in which beam search works to generate 3 distractors.

5.2.3 Effect of Cosine Term

The goal of the cosine similarity term is to avoid distractors similar to the correct answer. As a qualitative comparison we looked for examples in the predictions of T5-small where the similarity of the answer and the prediction is high according to BLEU scores (BLEU 3 greater or equal to 10). Then check how the predictions look like in the model with modified loss function. We found some cases where the model with modified loss function successfully avoid a prediction similar to the correct answer as shown in Figure 5.

6 Discussion

6.1 Dataset Issues

Although we opted to use the preprocessed dataset for our work, we firmly believe that with more resources (and hence, a larger version of T5 and more epochs), using the original RACE dataset will return better results. Not only did the preprocessed dataset contain half the amount of distractors from the original set, but it also contained inconsistencies in the way distractors were formatted. The testing set contained a single ground-truth distractor per question, whereas the validation set had three ground truth distractors per question. This formatting difference complicated model evaluation across datasets, which is why we evaluated all of our models on the validation set.

6.2 Memory Limits

One of the biggest obstacles we found in the project is lack of resources at hand. Even though the GPU's RAM had enough capacity to run T5-small and perform our experiments with some limitations, we were at a disadvantage compared to our established baselines that ran models of higher complexity than ours. We believe that larger versions of T5 by themselves could get a better performance than other models if we train them with larger batch sizes and tune its hyperparameters. The performance of those larger models could be further improved by adding the cosine term to the loss function.

6.3 Alternative Approach to Distractor Evaluation

The modifications that we performed to the loss function aimed to improve the quality of the distractors by addressing some problems of other approaches, like the similarity of the distractors to the correct answer that would make the solution ambiguous. However, those problems are not fully captured by the evaluation metric (BLEU) because it only measures how similar the predicted distractor is to the set of ground truth distractors.

6.3.1 Flaw in BLEU metric for Distractor Evaluation

In order to truly measure the contribution of our work in producing effective distractors, we believe that a novel evaluation metric should be conceived that accounts for other features that define the quality of a distractor. In Figures 6 and 7 of the Appendix, we present two examples of what we consider good distractors that have low BLEU-4 scores. Evidently, the generated distractor is different to the ground truth. However, these instances do show that a generated distractor with low overlap can nonetheless be ”clever” and make grammatical *and* contextual sense. Hence, it is not far-fetched to suggest that BLEU scores could be a misleading metric in evaluating distractors.

6.3.2 Towards Human Evaluation

In future investigations in DG, we suggest running a pipeline that includes BLEU scores as an evaluation metric. But we also strongly recommend performing a human evaluation survey to obtain another perspective on the quality of distractors generated. To this date, unfortunately, we are yet to find an agreed-upon framework that describes how such a survey should be conducted. Nonetheless, a few approaches suggested by the literature provide interesting ways to go about this. Namely, Lelkes et al. (2021), Gao et al. (2018) and Offerijns et al. (2020).

7 Conclusions

In this work, we introduced a novel approach for distraction generation using T5 with a new prefix and a modified loss function. Our experiments revealed that the additional cosine similarity term in the loss function helps improve performance. Next, we think that using a larger version of T5 will likely increase the quality of the distractors and the BLEU metric performance. Finally, we discuss the limitations of the BLEU metric as an assessor of distractors and briefly consider the merits of conducting a human-led survey to assess the quality of distractors.

References

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A bert-based distractor generation scheme with multi-tasking and negative answer training strategies.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2018. Generating distractors for reading comprehension questions from real examinations.

Ayako Hoshino and H. Nakagawa. 2007. Assisting cloze test making with a web application.

Shu Jiang and John Lee. 2017. Distractor generation for Chinese fill-in-the-blank items. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Copenhagen, Denmark. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Ádám D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. *CoRR*, abs/2102.09094.

Kaushal Kumar Maurya and Maunendra Sankar Desarkar. 2020. Learning to distract: A hierarchical multi-decoder network for automated generation of long distractors for multiple-choice questions for reading comprehension. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 1115–1124, New York, NY, USA. Association for Computing Machinery.

Jeroen Offerijns, Suzan Verberne, and Tessa Verhoef. 2020. Better distractions: Transformer-based distractor generation and multiple choice question filtering. *CoRR*, abs/2010.09598.

J. Pino and M. Eskénazi. 2009. Semi-automatic generation of cloze question distractors effect of students' l1. In *SLaTE*.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic distractor generation for multiple choice questions in standard tests.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer.

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

Siyu Ren and Kenny Q. Zhu. 2020. Knowledge-driven distractor generation for cloze-style multiple choice questions.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost.

Ex-Students' Association U. of Texas. 2001. Publications manual.

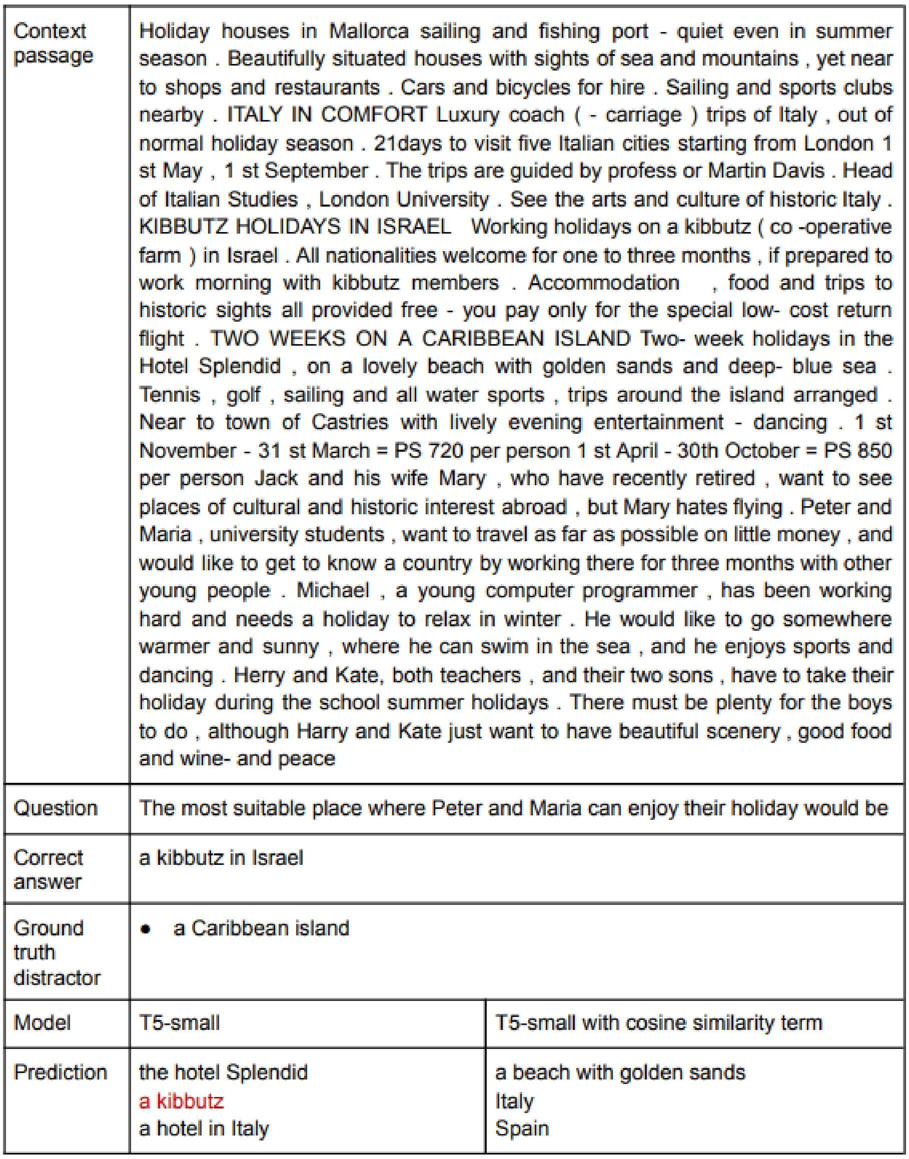
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2019. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension.

A Appendix

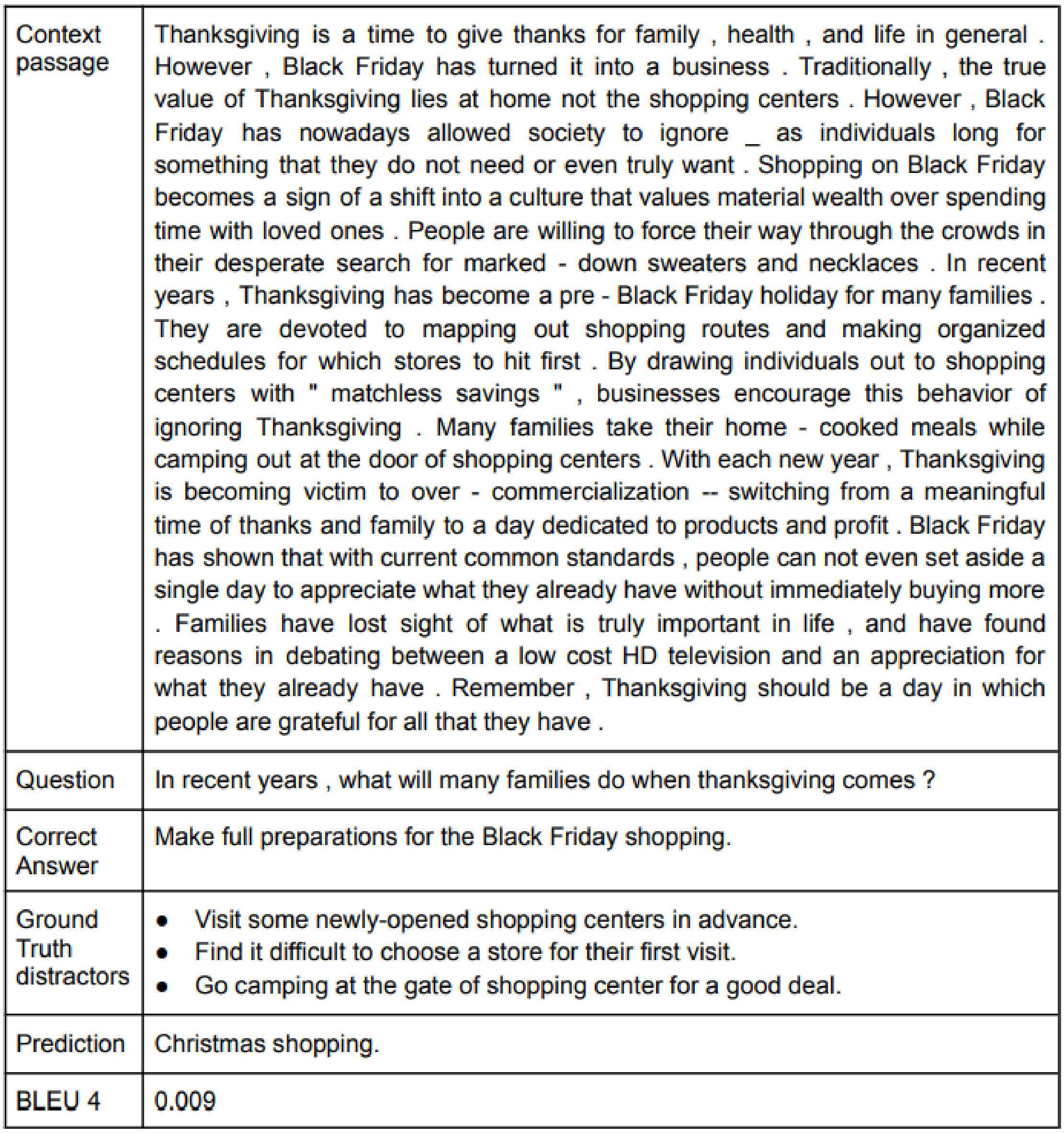


Context passage	Holiday houses in Mallorca sailing and fishing port - quiet even in summer season . Beautifully situated houses with sights of sea and mountains , yet near to shops and restaurants . Cars and bicycles for hire . Sailing and sports clubs nearby . ITALY IN COMFORT Luxury coach (- carriage) trips of Italy , out of normal holiday season . 21days to visit five Italian cities starting from London 1 st May , 1 st September . The trips are guided by profess or Martin Davis . Head of Italian Studies , London University . See the arts and culture of historic Italy . KIBBUTZ HOLIDAYS IN ISRAEL Working holidays on a kibbutz (co -operative farm) in Israel . All nationalities welcome for one to three months , if prepared to work morning with kibbutz members . Accommodation , food and trips to historic sights all provided free - you pay only for the special low- cost return flight . TWO WEEKS ON A CARIBBEAN ISLAND Two- week holidays in the Hotel Splendid , on a lovely beach with golden sands and deep- blue sea . Tennis , golf , sailing and all water sports , trips around the island arranged . Near to town of Castries with lively evening entertainment - dancing . 1 st November - 31 st March = PS 720 per person 1 st April - 30th October = PS 850 per person Jack and his wife Mary , who have recently retired , want to see places of cultural and historic interest abroad , but Mary hates flying . Peter and Maria , university students , want to travel as far as possible on little money , and would like to get to know a country by working there for three months with other young people . Michael , a young computer programmer , has been working hard and needs a holiday to relax in winter . He would like to go somewhere warmer and sunny , where he can swim in the sea , and he enjoys sports and dancing . Herry and Kate, both teachers , and their two sons , have to take their holiday during the school summer holidays . There must be plenty for the boys to do , although Harry and Kate just want to have beautiful scenery , good food and wine- and peace	
Question	The most suitable place where Peter and Maria can enjoy their holiday would be	
Correct answer	a kibbutz in Israel	
Ground truth distractor	• a Caribbean island	
Model	T5-small	T5-small with cosine similarity term
Prediction	the hotel Splendid a kibbutz a hotel in Italy	a beach with golden sands Italy Spain

Figure 5: Example distractor similar to the correct answer produced by T5-small. The distractor ”a kibbutz” can be considered as a correct answer.

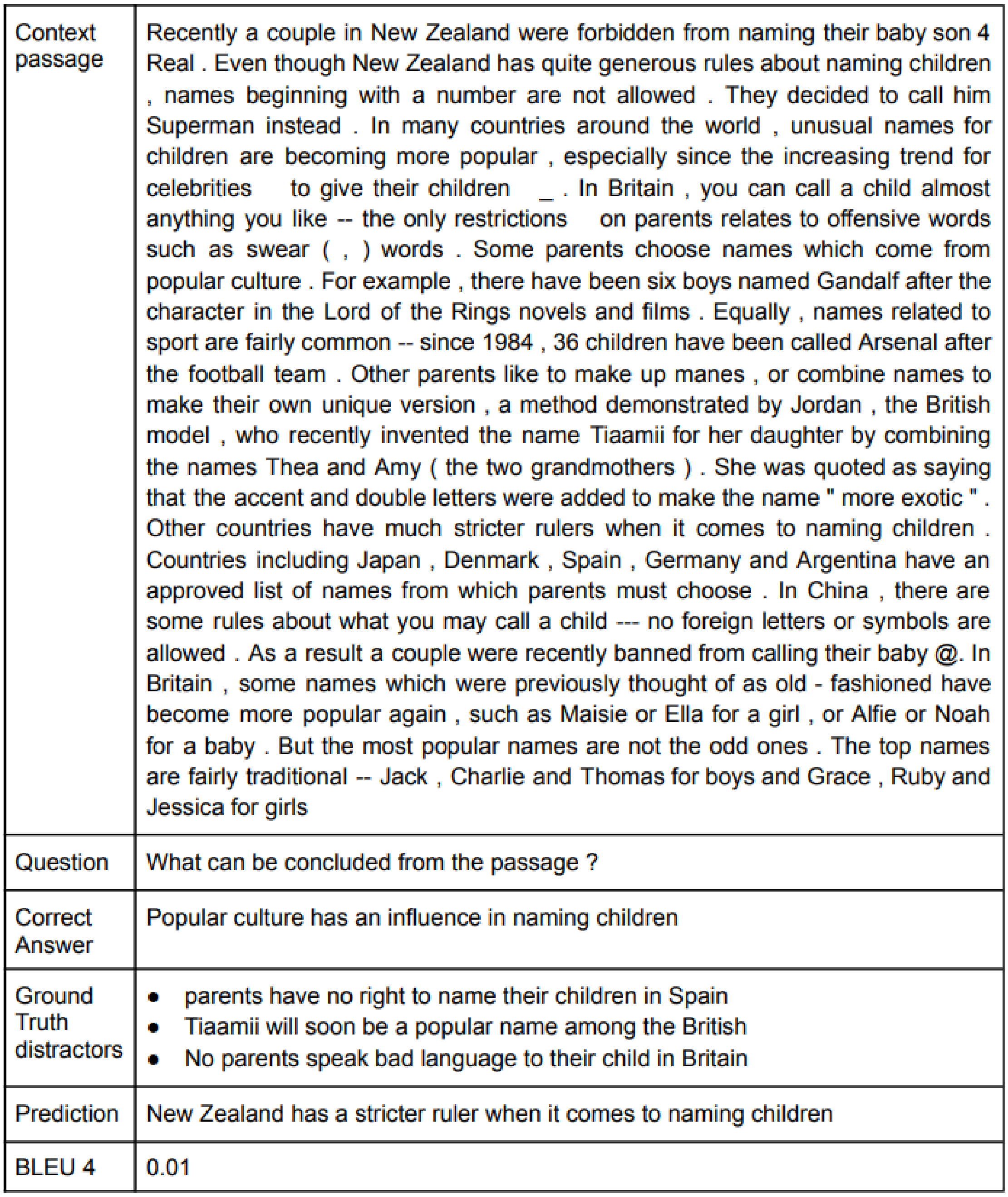
	1st distractor				2nd distractor				3rd distractor			
Lambda	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU-1	BLEU-2	BLEU-3	BLEU-4
0	24.34	8.03	4.50	3.45	21.36	7.66	4.40	3.17	20.47	6.73	3.67	**2.84**
0.1	23.03	8.29	4.82	3.45	**25.13**	**8.62**	**4.75**	**3.51**	**21.24**	**7.09**	**3.81**	**2.84**
0.2	23.73	8.38	4.77	3.41	24.29	8.07	4.48	3.33	20.10	6.54	3.53	2.61
0.3	**25.45**	9.06	5.13	3.73	24.86	8.13	4.37	3.23	18.41	5.75	3.04	2.24
0.4	25.36	**9.09**	**5.19**	**3.86**	24.69	8.28	4.66	3.44	19.52	6.34	3.35	2.58

Table 6: Scores for 3 distractors by lambda



Context passage	Thanksgiving is a time to give thanks for family , health , and life in general . However , Black Friday has turned it into a business . Traditionally , the true value of Thanksgiving lies at home not the shopping centers . However , Black Friday has nowadays allowed society to ignore _ as individuals long for something that they do not need or even truly want . Shopping on Black Friday becomes a sign of a shift into a culture that values material wealth over spending time with loved ones . People are willing to force their way through the crowds in their desperate search for marked - down sweaters and necklaces . In recent years , Thanksgiving has become a pre - Black Friday holiday for many families . They are devoted to mapping out shopping routes and making organized schedules for which stores to hit first . By drawing individuals out to shopping centers with " matchless savings " , businesses encourage this behavior of ignoring Thanksgiving . Many families take their home - cooked meals while camping out at the door of shopping centers . With each new year , Thanksgiving is becoming victim to over - commercialization -- switching from a meaningful time of thanks and family to a day dedicated to products and profit . Black Friday has shown that with current common standards , people can not even set aside a single day to appreciate what they already have without immediately buying more . Families have lost sight of what is truly important in life , and have found reasons in debating between a low cost HD television and an appreciation for what they already have . Remember , Thanksgiving should be a day in which people are grateful for all that they have .
Question	In recent years , what will many families do when thanksgiving comes ?
Correct Answer	Make full preparations for the Black Friday shopping.
Ground Truth distractors	● Visit some newly-opened shopping centers in advance. ● Find it difficult to choose a store for their first visit. ● Go camping at the gate of shopping center for a good deal.
Prediction	Christmas shopping.
BLEU 4	0.009

Figure 6: Example 1 of good quality distractor with low BLEU-4 score



Context passage	Recently a couple in New Zealand were forbidden from naming their baby son 4 Real . Even though New Zealand has quite generous rules about naming children , names beginning with a number are not allowed . They decided to call him Superman instead . In many countries around the world , unusual names for children are becoming more popular , especially since the increasing trend for celebrities to give their children _ . In Britain , you can call a child almost anything you like -- the only restrictions on parents relates to offensive words such as swear (,) words . Some parents choose names which come from popular culture . For example , there have been six boys named Gandalf after the character in the Lord of the Rings novels and films . Equally , names related to sport are fairly common -- since 1984 , 36 children have been called Arsenal after the football team . Other parents like to make up manes , or combine names to make their own unique version , a method demonstrated by Jordan , the British model , who recently invented the name Tiaamii for her daughter by combining the names Thea and Amy (the two grandmothers) . She was quoted as saying that the accent and double letters were added to make the name " more exotic " . Other countries have much stricter rulers when it comes to naming children . Countries including Japan , Denmark , Spain , Germany and Argentina have an approved list of names from which parents must choose . In China , there are some rules about what you may call a child --- no foreign letters or symbols are allowed . As a result a couple were recently banned from calling their baby @. In Britain , some names which were previously thought of as old - fashioned have become more popular again , such as Maisie or Ella for a girl , or Alfie or Noah for a baby . But the most popular names are not the odd ones . The top names are fairly traditional -- Jack , Charlie and Thomas for boys and Grace , Ruby and Jessica for girls
Question	What can be concluded from the passage ?
Correct Answer	Popular culture has an influence in naming children
Ground Truth distractors	● parents have no right to name their children in Spain ● Tiaamii will soon be a popular name among the British ● No parents speak bad language to their child in Britain
Prediction	New Zealand has a stricter ruler when it comes to naming children
BLEU 4	0.01

Figure 7: Example 2 of good quality distractor with low BLEU-4 score

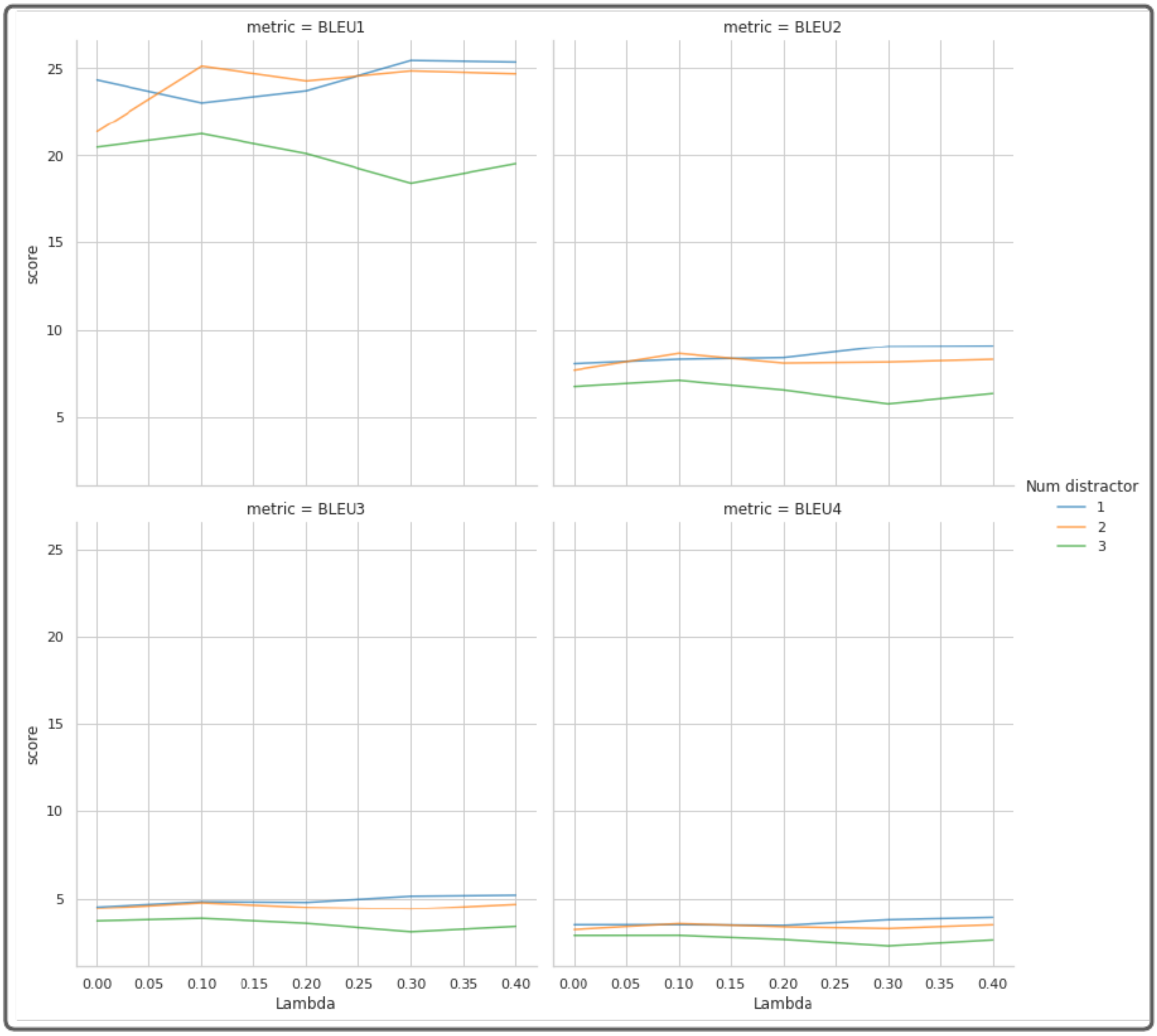


Figure 8: Performance of 3 distractors different values of lambda