Varifocal Question Generation for Fact-checking

Nedjma Ousidhoum* **Zhangdie Yuan*** **Andreas Vlachos**
Department of Computer Science and Technology
University of Cambridge
ndo24,zy317,av308@cam.ac.uk

Abstract

Fact-checking requires retrieving evidence related to a claim under investigation. The task can be formulated as question generation based on a claim, followed by question answering. However, recent question generation approaches assume that the answer is known and typically contained in a passage given as input, whereas such passages are what is being sought when verifying a claim. In this paper, we present *Varifocal*, a method that generates questions based on different focal points within a given claim, i.e. different spans of the claim and its metadata, such as its source and date. Our method outperforms previous work on a fact-checking question generation dataset on a wide range of automatic evaluation metrics. These results are corroborated by our manual evaluation, which indicates that our method generates more relevant and informative questions. We further demonstrate the potential of focal points in generating sets of clarification questions for product descriptions.

1 Introduction

The growing amount of information online and its impact have increased the need for fact-checking, i.e. judging whether a claim is true or false. To determine the truthfulness of a claim, fact-checkers need to answer questions related to the claim, world knowledge within its time frame, local politics, etc. (Graves, 2017). Using questions and answers has also been shown to be an effective way of conveying fact-checks. For instance, Altay et al. (2021) found that presenting information related to COVID-19 as answers to questions improved attitudes towards vaccination more than merely presenting the relevant facts.

As professional fact-checkers can spend a day to verify a single claim depending on its complexity (Adair et al., 2017; Hassan et al., 2017), there has been a growing focus on how to accelerate the fact-checking process via automation (Cohen et al., 2011; Graves and Anderson, 2020). Fan et al. (2020) showed that generating questions and answering them reduces the time spent on verification by approximately 20%. Fact verification questions tackle information that is missing from the claim, which renders the generation task challenging yet useful for assisting professionals.

Previous work on question generation assumes that the answer is known, typically contained in a passage given in the input (Rajpurkar et al., 2016; Duan et al., 2017; Wang et al., 2017). Such passages, though, are what is being sought when fact-checking a claim. The only exception is recent work on clarification questions (Rao and Daumé, 2019; Majumder et al., 2021). However, work in this area examines a specific narrow domain where a limited number of questions can be asked, e.g. questions related to product descriptions in the Amazon dataset of McAuley and Yang (2016), or dialogues in the Ubuntu dataset (Lowe et al., 2015), which is not the case in fact-checking. Fact-checking questions are more diverse and may rely on the experience and intuition of the fact-checker as they aim to scrutinize every piece of information within the claim. They can be as generic as questions about general definitions or as specific as those tackling details about a one-time event.

In this paper, we propose an approach that generates questions for claim verification, which we name *Varifocal*. It uses focal points from different spans of the claim as well as its metadata, i.e. its source, and its date. Each focal point guides the generator to question a different part of the claim; e.g. in Figure 1 when *"Miss Universe Guyana 2017"* is used as the focal point, the question generated is about who she is.

We evaluate our approach on the QABriefs dataset introduced by Fan et al. (2020) using a wide range of automatic metrics, and show that

*Equal contribution.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2532 - 2544

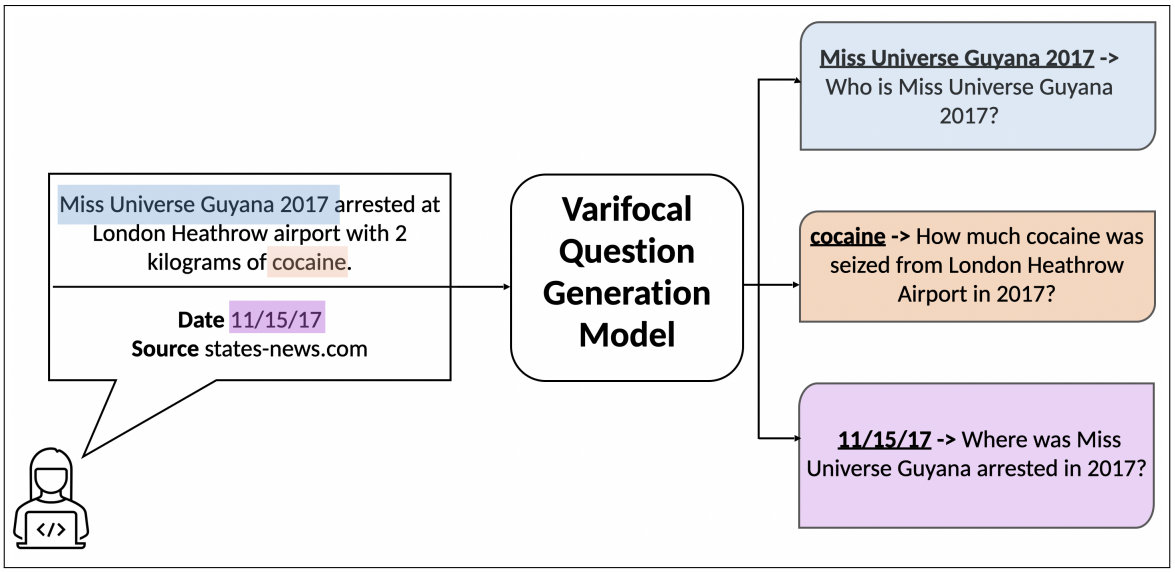

Figure 1: The architecture of Varifocal. We use a dependency parser to extract the different focal points, i.e. spans, then generate questions based on them. We rank the generated questions using a re-ranker and return the top n questions. The example in the figure was generated by our system. We show three highlighted focal points along with the (output) questions they led to.

Varifocal performs the best among the different systems considered. In addition, we conduct a human evaluation on questions generated by four different systems and gold standard questions based on four criteria: a) intelligibility, b) clarity, c) relevance, and d) informativeness. The results show that Varifocal generates more intelligible, clear, relevant, and informative questions than the other systems, corroborating the results of the automatic evaluation.[*]

Finally, we apply Varifocal to generating sets of clarification questions on Amazon product descriptions (McAuley and Yang, 2016), where it shows competitive performance against methods that generate single questions while having other ones in the set as part of the input (Majumder et al., 2021).

2 Related Work

The main piece of previous work on question generation for fact-checking is by Fan et al. (2020). They proposed the QABriefs dataset, which consists of claims with manually annotated question-answer pairs containing additional information about the claims (e.g. the exact definition of a term, the content of a bill, details about a political statement or a vote). The QABriefs dataset contained questions asked by crowd-workers, who had to read both the claim and its fact-checking article. Fan et al. (2020) presented the QABriefer model, which generates a set of questions conditioned on the claim, searches the web for evidence and retrieves answers. However, they evaluated the questions generated only using BLEU scores without conducting a human evaluation. More recently, Yang et al. (2021) addressed the problem of explainability in fact-checking through question answering using the Fool Me Twice corpus (Eisenschlos et al., 2021). They generated questions from the claim, retrieved answers from the evidence, and compared them to the generated ones. Yet, they did not evaluate their question generation process, and assumed that the evidence is given as input to generate the questions, which is unrealistic since the questions are typically used for evidence retrieval.

Other related work includes Saeidi et al. (2018) who introduced a dataset containing 32k instances of real-world policies, crowd-sourced fictional life scenarios, and dialogues in order to reach a final yes/no answer. The policies were given as input and were explicitly stating what information needed to be asked for, and the questions had to have a yes or no answer. Neither of these hold in fact-checking, where questions are not usually answered by yes or no, and the information to be searched for is not known in advance. More recently, Majumder et al. (2021) presented a method to generate clarification questions. They built a two-stage framework that identifies missing information using the notions

[*]Our code is available on https://github.com/nedjmaou/Varifocal_Fact_Checking_QG

of *global* and *local schemas*. The *global schema* was built using filtered key phrases extracted from contexts that were part of the same class of the data, e.g. a class of similar products in the Amazon data (McAuley and Yang, 2016) and similar dialogues in the Ubuntu dataset (Lowe et al., 2015), whereas the *local schema* was built using one given context, and they defined the missing information as the difference between the global and the local schema. The extraction of comparable schemas across different contexts was possible due to the repetitive nature of the datasets considered, e.g. the descriptions of products of the same type such as laptops allow the prediction of potentially missing properties which need clarification, in contrast to fact-checking claims which are less repetitive.

The standard sequence-to-sequence architecture (Sutskever et al., 2014) is typically used in question generation approaches (Du et al., 2017; Zhou et al., 2017). Although answer-aware approaches allow for the generation of multiple questions conditioned on the same passage (Sun et al., 2018), providing the answer during inference is not possible in fact-checking since one would typically ask questions about what is missing from the claim. Other work includes question generation for question answering (Duan et al., 2017), question generation for educational purposes (Heilman and Smith, 2010), and poll question generation from social media posts (Lu et al., 2021). Furthermore, Hosking and Riedel (2019) evaluated rewards in question generation, showed that they did not correlate with human judgments, and explained why rewards did not help when using reinforcement learning.

Commonly used evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) fall short at correlating with human judgments when evaluating the quality of automatically generated questions (Liu et al., 2016; Sultan et al., 2020; Nema and Khapra, 2018). Majumder et al. (2021) carried a human evaluation based on fluency, relevance, whether the question dealt with missing information, and usefulness. In addition, Cheng et al. (2021) proposed to assess the quality of automatically generated questions based on whether they were well-formed, concise, answerable, and answer-matching. Similarly, we conduct a human evaluation of the generated questions adapted to fact-checking.

3 Varifocal Question Generation

In this section, we describe *Varifocal*, an approach that generates multiple questions per claim based on its different aspects, which correspond to textual spans that we call *focal points*.

Varifocal consists of three components: (1) a focal point extractor, (2) a question generator that generates a question for each focal point, and (3) a re-ranker that ranks the generated questions, removes duplicates and promotes questions that are more likely to match the gold standard ones.

3.1 Focal Point Extraction

We consider two types of focal points: contiguous spans from the claim and metadata elements. For the former, we consider all the subtrees of its syntactic parse tree, thus obtaining more coherent phrases than if we extracted randomly selected n-grams. In addition, the metadata, which includes (1) the source of the claim or the name of the speaker, and (2) the date when the claim was made, can be useful in question generation for fact-checking. As shown in Figure 1, having access to the date of the claim helped the model generate a precise question, i.e. Where was Miss Universe Guyana arrested *in 2017*?. As the metadata is not part of the claim, we incorporate it using a template. For instance, we combined the claim and metadata of the example shown in Figure 1 as follows: *statenews.com reported on 11/15/17 that Miss Universe Guyana 2017 was arrested at London Heathrow airport with 2 kilograms of cocaine*.

3.2 Question Generation

This component takes a claim and its focal points as input and generates a set of questions. Given a claim c, the set of all focal points is denoted as F, where each focal point $f_i \in F$ is a span in the claim c and its metadata, such as $f_i = [w_s, ..., w_e]$ where s and e mark the start and the end of the span, respectively. Then, for each focal point f_i, the model generates autoregressively a question $\hat{q}_i$ of n words, as follows:

$$p(\hat{q}_i|c, f_i) = \prod_{k=1}^{n} p(\hat{q}_{i_{[k]}}|\hat{q}_{i_{[0:k-1]}}, [\tilde{c}; \tilde{f}_i]) \quad (1)$$

$[\tilde{c}; \tilde{f}_i]$ is the transformer-based encoding of c concatenated to f_i. The question generation component in Varifocal is similar to the answer-aware sequence-to-sequence model (Sun et al., 2018).

But, the generator is trained to use focal points instead of answers to the questions that need to be generated. While focal points act similarly to prompts (Radford et al., 2018; Brown et al., 2020; Liu et al., 2021), we typically have multiple focal points that are different for each claim, depending on the complexity of its parse tree, as opposed to a small number of fixed prompts, such as one per label in classification tasks.

3.3 Re-ranking

After all question candidates are generated (one per focal point), the re-ranker removes duplicated questions in addition to almost identical ones by setting a BLEU score threshold to 0.8.

The re-ranker then scores the remaining questions using a regression model, which assigns a real number score to each candidate. The more similar the candidate is expected to be to one of the gold questions, the higher score it should receive.

3.4 Training

To train the question generation model, we need focal points paired with the questions that they led to generate. However, most question generation datasets have questions paired with answers instead of focal points. Therefore, we use cosine similarity to match the extracted focal points $f_i \in F$ with the gold answers during training. I.e. we calculate $sim(emb(f_i), emb(a_j))$: the cosine similarity between the embeddings of f_i and the gold answer a_j. Following this, we greedily match each answer (and associated question) with the highest scoring focal point. We then remove the latter from the set of focal points available for matching. In our experiments, we generate the embeddings $emb(f_i)$ and $emb(a_j)$ using Sentence-BERT (Reimers and Gurevych, 2019).

We train the re-ranker on a holdout split, i.e. a small portion of the data that we did not use to build the generator. Given a claim c, the re-ranker g is trained to predict the similarity score that a question would have with the best matching question from the gold standard. Therefore, it considers the maximum sentence similarity of each of the generated questions $\hat{q}_i$ and the gold standard ones $q_j \in \mathcal{Q}$, with $\mathcal{Q}$ the set of gold questions associated to c. The objective function is expressed as follows:

$$L(D, \theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2)$$

where $y_i = \max_{q_j \in \mathcal{Q}} sim(emb(\hat{q}_i), emb(q_j))$, $\hat{y}_i = g(\theta, \hat{q}_i)$ is the score predicted by the re-ranker, D refers to the training data, $n = |D|$, and sim calculates the cosine similarity between the sentence embeddings $emb(\hat{q}_i)$ and $emb(q_j)$ of $\hat{q}_i$ and q_j, respectively. We use Sentence-BERT (Reimers and Gurevych, 2019) to compute sim.

4 Experimental Setup

4.1 Data

We train our question generation system on the QABriefs dataset, which contains 7,535 claims with 21,168 questions. We change the splits used by Fan et al. (2020) to ensure that all the claims in the test set contain metadata. Our training set contains 5,228 claims associated with 14,371 questions, the validation set has 653 claims associated with 1,958 questions, and the test set is composed of 653 claims associated with 1,952 questions. We also reserve a further holdout split of 999 claims for training the re-ranker. In our experiments, we use the SpaCy dependency parser (Honnibal and Montani, 2017) to parse the claims and extract the focal points.

4.2 Models

In our experiments, we train the following models.

- **BART** For each claim, we use BART (Lewis et al., 2019) to generate a set of questions separated by a separation token *[SEP]*. This is a replication of the QABriefer model reported by Fan et al. (2020).
- **SQuAD** We train an answer-aware question generation model (Sun et al., 2018) on the SQuAD dataset (Rajpurkar et al., 2018) without further fine-tuning on the QABriefs dataset. However, we use focal points instead of the answers. We extract the focal points using the method described in Section 3.4.
- **Varifocal** We pretrain the model on SQuAD and fine-tune it on the QABriefs dataset. The models take tuples of the form (c, f_i, q_i) as input.
- **Varifocal+Meta** We pretrain the model on SQuAD and fine-tune it on the QABriefs dataset with the metadata associated with each claim.

We generate different datasets for training a re-ranker for each of the models considered, i.e. SQuAD, Varifocal, and Varifocal+Meta, following Section 3.4.

4.3 Automatic Evaluation

We evaluate our question generation models using:

- **BLEU** (Papineni et al., 2002) which evaluates n-gram precision (used $n = 2$ and $n = 4$),
- **chrF** character n-gram F-score (Popović, 2015),
- **METEOR** (Banerjee and Lavie, 2005) which is based on unigram precision and recall as well as stemming and synonymy matching for similarity,
- **ROUGE** (Lin, 2004) which evaluates n-gram overlap (used $n = 1$ $n = 2$),
- **ROUGE-L** which uses the Longest Common Subsequence statistic,
- **TER** (Snover et al., 2006) which is an error rate based on edit distance.

4.4 Human Evaluation

We conduct a human evaluation of the generated questions based on a) intelligibility, b) clarity, c) relevance, and d) informativeness. The raters assign a 0/1 score to the intelligibility, clarity and relevance of the question, and a score from 0 to 3 to its informativeness. We define these criteria in the context of fact-checking as follows.

Intelligibility The question should be fluent, and as long as it is understandable, it does not have to be perfectly grammatical. The intelligibility of a question should be judged without looking at the claim. This criterion is similar to the *fluency* criterion presented by Majumder et al. (2021) and to the *good form* criterion proposed by Cheng et al. (2021).

Clarity Questions should be clear enough to be answered confidently using a search engine. Hence, a clear question should not be too broad and should include some necessary details, such as the date and the name of the speaker, etc. The question remains clear if these details can be induced by looking at the claim.

Relevance A generated question is only relevant if it mentions entities related to the claim. The entities can either be mentioned in the claim or the metadata since we use the latter to train a question generation system.

Informativeness An informative question should return answers providing information that helps us judge the veracity of the claim. The informativeness of a fact-checking question depends on the nature of the claim. For instance, if the claim is a quote, a question that focuses on the person or entity who made the statement can be informative. On the other hand, if the claim is about an event, then an informative fact-checking question may be about the event itself. While questions should not directly ask whether the claim is true or false, a yes/no question which is necessary to reach a verdict is informative. We use a 4-point Likert scale to assess the informativeness of a question. A question can be:

1. **uninformative** i.e. useless *(score = 0)*,
2. **weakly informative** i.e. unlikely to be helpful but we do not mind having it generated by the system *(score = 1)*,
3. **potentially informative** or somewhat useful, i.e. a question that could be helpful depending on the context of the claim. For instance, a question whose answer is in the claim can be informative if it is worth verifying *(score = 2)*,
4. **informative** i.e. a question to which the answer is crucial for the fact-check *(score = 3)*.

An informative question is intelligible, clear, and relevant. Informativeness differs from the *missing information* criterion defined by Majumder et al. (2021) in that it is not defined against a predefined schema such as a product description.

We asked three volunteers to assess the quality of the questions. The raters are researchers in NLP (not involved in the paper) working in an English-speaking institution. They were assigned ten questions per claim, i.e. two gold questions associated with the claim in the QABriefs dataset and two questions generated by each of our models: SQuAD, BART, Varifocal and Varifocal+Meta. We hid the name of the systems which generated the questions and their ranking from the raters.

5 Results

5.1 Automatic Evaluation

The results presented in Table 1 show consistency with respect to which system performs the best despite the variation in their rankings. The Varifocal systems outperform BART and SQuAD by around 4 ROUGE-1 points, >4 METEOR points, and show lower error rates based on TER. Interestingly, SQuAD slightly outperforms BART despite not being trained on the QABriefs dataset. However, SQuAD uses focal points to generate questions, which indicates their potential usefulness.

System	BLEU-2	BLEU-4	chrF	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	TER
BART	25.63	11.83	37.25	31.57	33.66	14.02	33.34	0.804
wh-BART	21.88	10.54	40.1	33.62	29.97	13.19	29.46	0.8697
SQuAD	25.85	11.95	38.60	30.67	32.70	13.63	31.84	0.809
Varifocal	29.98	15.17	41.12	34.84	37.27	17.64	36.77	**0.755**
Varifocal+Meta	**30.18**	**15.54**	**43.17**	**37.02**	**38.19**	**18.37**	**37.59**	0.764

Table 1: Automatic evaluation results on the QABriefs test set. For all scores higher is better except for TER.

Avg	I	C	R	Info
Gold	**0.97**	0.91	0.79	1.72
SQuAD	0.83	0.84	0.77	1.91
BART	0.85	0.76	0.67	1.49
Varifocal	**0.97**	**0.94**	**0.93**	**2.33**
Varifocal+Meta	0.93	0.91	0.89	2.10

Table 2: Average of intelligibility (I), clarity (C), relevance (R), and informativeness scores per system based on our human evaluation.

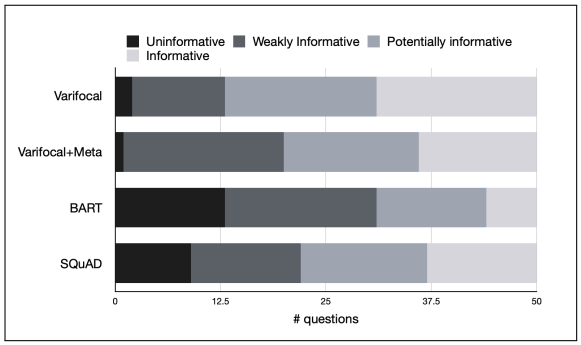

Figure 2: The distribution of informativeness scores across the different systems. Brighter means better.

We notice a minor difference between Varifocal and Varifocal+Meta except for TER, where Varifocal (without metadata) performs slightly better. To further assess the potential of focal points, we experimented with the BART question generator . We forced a BART model (wh-BART) to initiate the generated questions with the most common question words in the QABriefs dataset, typically wh-words such as *What*, *Why*, *How*, etc. (see Figure 4 in Fan et al. (2020)). While this resulted in scores comparable to those of BART and SQuAD, these were always lower than the scores achieved by Varifocal, further demonstrating that focal points provide guidance beyond the question type.

5.2 Human Evaluation

The raters evaluated a total of 250 questions generated for 25 different claims. As they assigned Boolean values to intelligibility (I), clarity (C), and relevance (R), similar distributions with minor disagreements led to low Fleiss-κ and Krippendorff-α scores. The Fleiss-κ scores are 0.48 for intelligibility, 0.43 for clarity, 0.32 for relevance and 0.26 for informativeness. However, as reported in previous work, chance-adjusted scores can be low despite a high agreement due to their inappropriateness when assessing variables with imbalanced marginal distributions (Randolph, 2005; Falotico and Quatto, 2015; Yannakoudakis and Cummins, 2015; Matheson, 2019). Therefore, we have computed the free marginal multi-rater kappa scores (Randolph, 2005), which are equal to 0.88 for intelligibility, 0.74 for clarity, 0.58 for relevance and 0.32 for informativeness.

Unsurprisingly, intelligibility is the least subjective criterion and is therefore the one on which the raters disagree the less, followed by clarity and relevance. On the other hand, we report a relatively low agreement on the informativeness criterion. Table 3 shows examples of questions generated by the different models and how they were rated.

The results of our human evaluation corroborate the scores of the automatic metrics presented in Section 5.1. Overall, Varifocal and Varifocal+Meta generate the best questions on average based on the four criteria. Moreover, they seem to generate more relevant and informative questions even when compared to the gold ones. In fact, some gold questions consider non-trivial prior knowledge about a claim and, thus, are sometimes annotated as irrelevant, i.e. they mention entities and events that are not part of the claim and the metadata. This is due to the fact that when building the QABriefs dataset, the annotators had to read the fact-checking article associated with the claim. As a result, the annotators have sometimes assumed the veracity/falsehood of a claim when asking questions and mentioned entities and events that only appeared in the fact-checking article. For instance, the gold question of claim 3 shown in Table 3 refers to an act (overseeing elections) not mentioned in the claim. We observe the most considerable difference in the high informativeness range (score=3), as shown

		Scores			
Claim 1	We have trade deficits with almost every country. *(Donald Trump, 7/28/17)*	**I**	**C**	**R**	**Info**
BART	What was the trade deficit with China as Secretary of State?	0	0	0	0
Varifocal	*trade* ⟶ What is the trade deficit with China?	1	1	1	1.67
BART	What are the trade deficits with almost every country?	0.33	0	0	0
Claim 2	Black home ownership just hit the highest level it has ever been in the history of our country. *(Donald Trump, 12/8/17)*	**I**	**C**	**R**	**Info**
Meta+Varifocal	*of our country* ⟶ What is the highest black home ownership in the history of the US?	1	1	1	3
Gold	How true is it that black ownership is at its highest level ever?	1	1	1	1.33
BART	What is the average black home price in the U.S.?	0.33	0.33	0.33	1
Claim 3	#Broward elections department has a history of violating the law. *(Marco Rubio, 11/8/18)*	**I**	**C**	**R**	**Info**
Meta+Varifocal	*#Broward elections department* ⟶ Did the Broward elections department violate the law?	1	1	1	3
Gold	How will Rubio oversee that the election is done correctly?	1	1	0.67	0.67
Meta+Varifocal	*1/8/18* ⟶ When did the Broward elections department violate the law?	1	1	1	2
BART	What is the history of violating the law?	1	1	0	0

Table 3: Examples of annotated questions that are either generated by one of the different systems or present in the gold set of questions. We show examples with different averaged human ratings for 1) intelligibility (I), 2) clarity (C), 3) relevance (R) and 4) informativeness (Info) scores. The focal points associated with the questions generated by the Varifocal systems are also presented (in the text preceding the arrow).

in Figure 2. The BART model, i.e. the approach used in the QABriefer, fails noticeably to generate fact-checking questions of high informativeness. A relatively large percentage of BART questions were found to be ineffective for fact-checking since an informative question needs to be intelligible, clear, and relevant. Hence, the unclear questions were uninformative by default per our guidelines.

5.3 Evaluation on the Amazon Dataset

To evaluate the potential of Varifocal and focal points outside fact-checking, we used the Amazon dataset of McAuley and Yang (2016). We trained our system (pretrained on SQuAD) on Amazon product descriptions only. Following Section 3.4, we extracted the focal points using SpaCy (Honnibal and Montani, 2017), trained a re-ranker, and then generated and ranked multiple questions per product. We achieved a BLEU-4 score of 13.2 using only the focal points as guidance. In contrast to the reported experimental setup by Majumder et al. (2021), who used previously asked questions as part of the input, effectively predicting the missing question in a set, we generate all questions given the product descriptions only. Their best model was trained on product descriptions, questions on the product, and previously asked questions about related products required to model missing information based on the notion of schema, i.e. *global schema - local schema*. This model achieves a BLEU-4 score of 18.55 and conditions BART on a usefulness classifier during decoding using Plug and Play Language Models (PPLMs) (Dathathri et al., 2020).

Nevertheless, among their baselines, they also trained a Transformer model that achieved a 12.89 BLEU score. This baseline was outperformed by Varifocal, although our system generated each question given the product description without access to the other questions asked about the product.

In conclusion, although modelling missing information using a schema can be useful, it can only be applied to a narrow domain, such as similar product descriptions. It is also worth noting that generating fact-checking questions can be harder than generating clarification ones. For instance, for the claim *"#Broward elections department has a history of violating the law."* shown in Table 3, the question *"What is the history of violating the law?"* can be considered a clarification question, but not a good fact-checking one.

Claim	Says Donald Trump promised "the mass deportation of Latino families."
Question	Did Trump promise mass deportation of Latino families? $\longrightarrow tag(f_i) = pobj$
Claim	Says NRA head Wayne LaPierre said, We believe in absolutely gun-free, zero-tolerance, totally safe schools. That means no guns in America's schools. Period.
Question	What is Wayne LaPierre's stance on guns in schools? $\longrightarrow tag(f_i) = nsubj$
Claim	Bay Area liberals have given more to Jon Ossoff's campaign than people in Georgia.
Question	How much money do people in Georgia give to Jon Ossoff? $\longrightarrow tag(f_i) = prep$

Table 4: Examples of questions with informativeness scores = 3 generated for focal points with tags $\in \{nsubj, pobj, prep\}$. The focal points are underlined in their respective claims.

	BLEU-4	**ROUGE-2**	**METEOR**
NE	15.44	35.15	17.48
Non NE	**16.13**	**36.88**	**18.33**

Table 5: BLEU-4, ROUGE-2 and METEOR scores achieved for focal points that are named entities (NE) only vs. other focal points (Non NE).

Focal points	**Avg(Informativeness)**
All	2.33
$tag(f_i)$=nsubj	2.18
$tag(f_i)$=dobj	2.58
$tag(f_i)$=pobj	2.49
$tag(f_i)$=prep	**2.63**
$tag(f_i)$=compound	2.37

Table 6: Average informativeness scores of the questions generated for focal points $f_i \in F$ whose tags are nsubj, dobj, pobj, prep and compound, respectively. *All* refers to the average informativeness score of all the questions generated by Varifocal that were labeled by our raters.

6 Analysis

To investigate the quality of the different focal points and how they guide the question generation process (1) we extracted focal points that are named entities only and those that are not, then automatically evaluated the performance of Varifocal on each of these subsets; (2) we compared the average human rating scores of all focal points to the ones achieved by subsets of focal points referring to specific syntactic tags. We show examples of questions generated for focal points with some of these tags in Table 4.

Table 5 shows that not considering named entities as focal points performs better than only considering named entities. This observation is especially relevant for fact-checking since named entities are often used in heuristics that generate questions (Lewis et al., 2021). On the other hand, this also proves that reducing the number of focal points does not necessarily affect the quality of the question generation.

When analysing the results, we observed that focal points whose syntactic roles are *nsubj*, *dobj*, *pobj*, *prep* and *compound* seem to lead to the best questions. We, therefore, examined their average informativeness scores according to the human raters. As shown in Table 6, the top 5 tags have an average informativeness score that is >2.

7 Limitations and Future Work

The Complexity of the Claims In our experiments on the QABriefs dataset, the maximum number of focal points was 175, as shown in Table 7 despite fact-checking claims being short in length and limited in terms of complexity. As the average number of focal points was 22 (or 28 when using metadata), we over-generated questions causing a fair amount of duplicates that needed to be removed by the re-ranker. However, as shown in Section 6, using a subset of focal points can be sufficient to achieve a comparable performance, which can help us reduce the running time of our method. By knowing in advance the types of focal points that would more likely lead to good questions, we can avoid extracting all the possible ones and reduce the complexity of our method.

Additional Evaluation In the future, besides evaluating the quality of individual questions in terms of intelligibility, clarity, relevance and informativeness, we intend to assess the quality of sets of questions. Metrics such as Distinct-2 (Li et al., 2016) used by Majumder et al. (2021) assess the lexical diversity of the sets whereas *semantic* diversity is the one that is critical for fact-checking (Sultan et al., 2020). The evaluation of a set of questions needs to take different dimensions into account and can depend on the nature of the claim for which we generate questions (e.g. a statement vs. an opinion). Furthermore, we plan to assess

	Min	Avg	Max
Varifocal	4	22	169
Varifocal+Meta	10	28	175

Table 7: Maximum, minimum, and average numbers of extracted focal points per system, i.e with and without using metadata.

questions considering the limitations of currently used search engines, i.e. whether the question is numerical, comparative, ambiguous, and whether its answer can, therefore, be easily fetched or not.

Knowledge Beyond the Claim One can also argue that only using the claim and the metadata to generate focal points and questions is insufficient for fact-checking. In the future, we propose to predict potential answers to the generated questions using concepts and entities related to different parts of the claim with the help of knowledge bases and search engines.

8 Conclusion

We introduced Varifocal, a question generation method for fact-checking that alleviates the absence of full context using focal points.

Varifocal, with and without metadata, improves the quality of automatically generated questions compared to other systems. It generates intelligible and clear questions, and sometimes questions that can be more relevant and informative than the gold standard ones. Moreover, when we used Varifocal to generate sets of clarification questions, it showed comparable results to those achieved by models that generate single questions while having additional ones as part of the input.

In the future, we will consider extensions using additional knowledge related to the different aspects of the claim and assess the overall quality of sets of questions.

9 Acknowledgements

We acknowledge Marzieh Saeidi for early contributions to this work. We thank Nicolas Patz and colleagues at Deutshe Welle, Andrea Parker, and Bryan Chen for evaluating previous versions of our system. We also thank Michael Schlichtkrull, Pietro Lesci, and Zhijiang Guo for their help and insightful feedback, as well as Afonso Mendes and anonymous reviewers and meta-reviewer for their comments on earlier versions of the manuscript.

Nedjma Ousidhoum is supported by the EU H2020 grant MONITIO (GA 965576). Zhangdie Yuan is supported by the ERC grant AVeriTeC (GA 865958). Andreas Vlachos is supported by both AVeriTeC and MONITIO.

References

Bill Adair, Chengkai Li, Jun Yang, and Cong Yu. 2017. Progress toward "the holy grail": The continued quest to automate fact-checking. In *Computation+ Journalism Symposium,(September)*.

Sacha Altay, Anne-Sophie Hacquin, Coralie Chevallier, and Hugo Mercier. 2021. Information delivered by a chatbot has a positive impact on covid-19 vaccines attitudes and intentions. *Journal of Experimental Psychology: Applied*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yi Cheng, Siyao Li, Bang Liu, Ruihui Zhao, Sujian Li, Chenghua Lin, and Yefeng Zheng. 2021. Guiding the growth: Difficulty-controllable question generation through step-by-step rewriting. In *Proceedings of ACL*, pages 5968–5978.

S Cohen, C Li, J Yang, and C Yu. 2011. Computational journalism: A call to arms to database researchers. In *5th Biennial Conference on Innovative Data Systems Research, CIDR*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation. In *Proceedings of ICLR*.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of ACL*.

Nan Duan, Duyu Tang, Peng Chen, and Ming Zhou. 2017. Question generation for question answering. In *Proceedings of EMNLP*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from wikipedia gamification. In *Proceedings of NAACL*.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*, 49(2):463–470.

Angela Fan, Aleksandra Piktus, Fabio Petroni, Guillaume Wenzek, Marzieh Saeidi, Andreas Vlachos, Antoine Bordes, and Sebastian Riedel. 2020. Generating fact checking briefs. In *Proceedings of EMNLP*, pages 7147–7161.

Lucas Graves. 2017. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, Culture & Critique*, 10(3):518–537.

Lucas Graves and Charles W Anderson. 2020. Discipline and promote: Building infrastructure and managing algorithms in a "structured journalism" project by professional fact-checking groups. *New Media & Society*, 22(2):342–360.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of KDD*.

Michael Heilman and Noah A. Smith. 2010. Good question! statistical ranking for question generation. In *Proceedings of NAACL-HLT*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings. *convolutional neural networks and incremental parsing*, 7(1).

Tom Hosking and Sebastian Riedel. 2019. Evaluating rewards for question generation models. In *Proceedings of NAACL*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of EMNLP*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of SIGDial*.

Zexin Lu, Keyang Ding, Yuji Zhang, Jing Li, Baolin Peng, and Lemao Liu. 2021. Engage the public: Poll question generation for social media posts. In *Proceedings of ACL*, pages 29–40.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian J. McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of NAACL*.

Granville J Matheson. 2019. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ*, 7:e6918.

Julian McAuley and Alex Yang. 2016. Addressing complex and subjective product-related queries with customer reviews. In *Proceedings of WWW*, pages 625–635.

Preksha Nema and Mitesh M Khapra. 2018. Towards a better metric for evaluating question generation systems. In *Proceedings of EMNLP*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings WMT*, pages 392–395.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of EMNLP*.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Sudha Rao and Hal Daumé. 2019. Answer-based adversarial training for generating clarification questions. In *Proceedings of NAACL-HLT*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of EMNLP*, pages 2087–2097.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231.

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 ACL*.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3930–3939.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of ACL*, pages 189–198.

Jing Yang, Didier Vega-Oliveros, Taís Seibt, and Anderson Rocha. 2021. Explainable fact-checking through question answering. *arXiv preprint arXiv:2110.05369*.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 662–671. Springer.

A Annotation Guidelines

We are evaluating a question generation system for fact-checking. First, we assess each question independently, then we evaluate the whole set of generated questions.

We assess each generated question based on the following criteria.

A.1 Intelligibility

The question should be fluent but does not have to be perfectly grammatical as long as it is understandable. The intelligibility of a question should be judged without looking at the claim.

Examples of intelligible vs. unintelligible questions

- **Unintelligible question** How many less do Florida's teachers pay? (*Incomprehensible.*)
- **Intelligible question** What made Rep. Paul Gosar to ask for the arrest of the illegal immigrants? (*Not perfectly grammatical but still intelligible.*)
- **Intelligible question** What is the average pay for Florida's teachers? (*Grammatical and intelligible.*)

A.2 Clarity

Questions should be precise enough to be answered confidently using a search engine regardless of the context. A clear question should not be too broad and should include all the necessary details, such as dates, and names of people/speaker, etc. If the details can be induced by looking at the claim, the question remains clear.

Examples of clear vs. unclear questions

- **Unclear question** What is the name of the state that New Jersey elects a Republican to the Senate? (*Unintelligible and unclear.*)
- **Unclear question** What policies violate federal law? (*Too broad.*)
- **Unclear question** What did the author of the bill say about the bill? (*Intelligible but unclear.*)
- **Clear question** What is the definition of a sanctuary city?
- **Clear question** What is the United Nations?

- **Clear question** What was the name of the law that separated children from adults entering America?

- **Claim** Apprehension rates at the southern border have plummeted since the 1980s and apprehensions of Mexicans specifically have reached their lowest point in nearly half a century.
  - **Clear question when looking at the claim (otherwise, unclear since the name of the country is not specified)** What was the apprehensions rate at the southern border in the 1980s?

A.3 Relevance

The generated questions should mention entities that are related to the claim. The entities can either be mentioned in the claim or in the metadata since we may use the latter to train a question generation system.

Examples of relevant vs. irrelevant questions

- **Claim** Miss Universe Guyana 2017 arrested at London Heathrow airport with 2 kilograms of cocaine.
  - **Irrelevant** Why would someone make up a fake news story about her hiding cocaine in coffee bags?
  - **Irrelevant** Why would someone make up a fake news story about her hiding cocaine in coffee bags?
  - **Relevant** Who was the Miss Universe Guyana 2017 arrested at London Heathrow airport with 2 kilograms of cocaine?
  - **Relevant** Who is Miss Universe Guyana 2017?
  - **Relevant** What is the name of the person arrested at London Heathrow airport?

A.4 Informativeness

An informative question should return answers that provide information about the claim in order to help us reach a verdict for its veracity. The informativeness of a fact-checking question will depend on the type of the claim. For instance, if the claim is a quote, a question which focuses on the person or entity who made the statement can be an informative one. On the other hand, if the claim focuses on the narration of a certain event, then an informative fact-checking question may focus on the event itself. A yes/no question which is useful to reach a verdict is considered to be informative. Questions, however, should not directly ask or imply that the claim is true or false. Finally, an informative question should not (indirectly) imply that the claim is true or false.

We use a 4-point Likert scale to assess the informativeness of a question. A question can be:

1. **uninformative** i.e. useless (0),
2. **weakly informative** i.e. unlikely to be helpful but we do not mind to have it generated by the system*(score = 1)* ,
3. **potentially informative** or somewhat useful, i.e. a question that could be helpful depending on the context of the claim. For instance, a question whose answer is in the claim can be informative if it is worth verifying *(score = 2)*,
4. **informative** i.e. crucial *(score = 3)*.

Examples questions scored according to their informativeness

- **Claim** Miss Universe Guyana 2017 arrested at London Heathrow airport with 2 kilograms of cocaine.
  - **Irrelevant and uninformative** Why would someone make up a fake news story about her hiding cocaine in coffee bags?
  - **Relevant and weakly informative** What is the name of the person arrested at London Heathrow airport?
  - **Relevant and potentially informative** Who is Miss Universe Guyana 2017?

- **Claim** You will learn more about Donald Trump by going down to the Federal Election Commission to see the financial disclosure form than by looking at tax returns.
  - **Relevant and uninformative** Where is Donald Trump's tax return?
  - **Relevant and weakly informative** How can you learn more about Donald Trump by looking at tax returns?

  - **Relevant and weakly informative** How can you learn more about Donald Trump by going down to the Federal Election Commission?
  - **Relevant and potentially informative** How much does Donald Trump donate to charity?
  - **Relevant and informative** What is Donald Trump's tax rate?
  - **Relevant and informative** What type of taxes does Donald Trump pay?

- **Claim** If Congress fails to act the Obama administration intends to give away control of the internet to an international body akin to the United Nations.
  - **Relevant and uninformative** What constitutes an international body?
  - **Relevant and uninformative** What would happen if Congress did not act?
  - **Relevant and weakly informative** Who has oversight over the internet in America?
  - **Relevant and weakly informative** What countries have officers involved in the Internet Corporation for Assigned Names and Numbers?
  - **Relevant and weakly informative** What is the United Nations?
  - **Relevant and potentially informative** What did the Obama administration intend to do with control of the internet?
  - **Relevant and informative question** What organization does the Obama administration want to give control of the internet to?

A.5 Prerequisites for the different criteria

1. Intelligibility should be judged without looking at the claim.
2. A clear question should be intelligible. When assessing the clarity, the claim can be checked for more details.
3. Relevance and informativeness should be annotated by looking at the claim.
4. A relevant question needs to be intelligible.
5. An informative question is intelligible, clear, and relevant.

B Implementation Details

B.1 Pre-processing

We remove duplicates by setting a 0.8 BLEU threshold so that extremely similar questions are removed before the re-ranking.

B.2 Models and Hyperparameters

We use the standard HuggingFace implementation of BART (https://huggingface.co/transformers/model_doc/bart.html) and BERT (https://huggingface.co/transformers/model_doc/bert.html).

B.2.1 Generator

BartForConditionalGeneration
Encoder layers 12
Encoder heads 16
Decoder layers 12
Decoder layers 16
Dimensionality 1024
Feed-forward layers dimensionality 4096
Activation function gelu
Dropout 0.1
Batch size 2 per device
Early stopping patience 10

B.2.2 Re-ranker

BertForSequenceClassification
Number of labels 1 (regression)
Encoder layers 12
Encoder heads 12
Dimensionality 768
Feed-forward layers dimensionality 3072
Dropout 0.1
Activation function gelu
Batch size 128 per device
Early stopping patience 10