A Zero-Resource Approach to Cross-Lingual Query-Focused Abstractive Summarization

Anonymous ACL submission

Abstract

We present a novel approach for cross-lingual query-focused abstractive summarization (QFAS) that leverages the translate-then-summarize paradigm. We approach cross-lingual QFAS as a zero-resource problem and introduce a framework to create a synthetic QFAS corpus from a standard summarization corpus using a novel query-generation strategy. Our model summarizes documents in foreign languages for which translation quality is poor. It learns not only to identify and condense salient information relevant to a query, but also to appropriately rephrase grammatical errors and disfluencies that may occur in the noisy translations. Our technique enhances a pre-trained encoder-decoder transformer by introducing query focus to the encoder. We show that our method for creating synthetic QFAS data leads to more robust models that not only achieve state-of-the-art performance on our corpus, but also perform better on out-of-distribution data as compared to prior work.

1 Introduction

Single document query-focused summarization (QFS) refers to the task of producing summaries that condense the salient information in a document that is pertinent to a query. This means that for the same document, different summaries can be produced depending on the input query. In the cross-lingual setting, the goal is to produce a summary in a target language, given a document in a source language and a query in the target language. In this paper, we focus on a configuration of this problem where the source document is in a foreign language while the target summary is in English.[1]

The current overload of digital content has made QFS an important task, enabling the quick consumption of information required by a user for a particular task. Since documents often have multi-faceted content, generating query-focused summaries relevant to people's interests and/or a particular task can be more useful than generic summaries. Cross-lingual summarization augments the benefits of QFS by enabling people to gain access to information written in languages that they do not understand.

The two main paradigms for cross-lingual summarization have been translate-then-summarize and summarize-then-translate (Wan et al., 2019). While summarize-then-translate might be computationally efficient since translation is done on reduced text, it can only be applied to high resource foreign languages where large summarization corpora are available (Ouyang et al., 2019). Since annotated translation data is more commonly available in larger scale than summarization data, the former paradigm is favorable since it is applicable to a broader class of foreign languages. In addition to this, even if the translations are of poor quality, the summarization model can leverage information redundancy to pick information from where it is translated more fluently. Errors from the summarization model are harder to recover from in the other paradigm. For these reasons, we adopt the translate-then-summarize approach.

One of the main concerns with this pipeline paradigm is the propagation of errors from machine translation (Zhu et al., 2019). This issue is particularly pronounced for lower resource foreign languages for which large-scale in-domain parallel translation corpora may not be available. Translation models trained on out-of-domain corpora (e.g., the Bible (Christodouloupoulos and Steedman, 2015) or EuroParl (Koehn, 2005)) may not transfer well. Models trained on small in-domain parallel corpora may not perform as well as those trained on large corpora in high resource languages. Thus error propagation is a glaring issue for extractive summarization models since the summary

[1] Though we focus on this configuration of the cross-lingual QFS problem, our approach could work for any pair of languages with available corpora.

contains disfluent sentences or phrases from the poorly translated document. However, abstraction can mitigate this issue by means of rewording.

The goal of our summarization model is thus two-fold: (a) to produce abstractive summaries that are relevant to a query; and (b) to improve potentially poor translations of foreign language documents provided as input.[2] The main contributions of this paper are:

- We introduce a new cross-lingual QFS corpus using a novel synthetic QFS corpus generation framework that generates more diverse and salient queries than contemporary approaches.
- We present a novel model architecture for cross-lingual query-focused abstractive summarization by augmenting pre-trained transformers, which, to our best knowledge, is the first attempt at the cross-lingual variant of the QFS task.
- Our summarization model outperforms prior work, based on both automatic metrics and human evaluation, on both our new corpus and an existing QFS corpus.

2 Dataset

While there are query-focused summarization datasets in the multi-document setting (Dang, 2006; Baumel et al., 2016; Pasunuru et al., 2021; Zhong et al., 2021), there is a lack of large annotated corpora for single-document QFS. This zero-resource setting can be handled by synthesizing a QFS corpus from pre-existing summarization corpora. In this work, we present a framework to generate a query-focused summarization corpus from a standard summarization corpus using a novel query-generation strategy. To build a summarization model that can handle poor translations during inference time, we follow Ouyang et al. (2019) and transform the generated QFS dataset to simulate this task using round-trip-translation to produce noisy (with translation disfluencies) documents paired with fluent summaries.

Our framework involves two components - (1) generation of QFS triples from the existing corpus; and (2) round-trip translation of the source document to introduce disfluencies. We synthetically generate a new cross-lingual QFS corpus using CNN-DailyMail (Vinyals et al., 2016; Nallapati et al., 2016) as our base corpus. Our dataset generation framework takes as input the $\{article, summary\}$ pairs in the CNN-DailyMail corpus to produce $\{article, query, summary\}$ triples. The generated triples have articles that are disfluent and summaries that only contain sentences that are relevant to the query.

[2] The code and data related to this paper can be found here upon paper acceptance

	Train	Validation	Test
Number of Instances	583,483	28,299	24,255
Number of Documents	284,435	13,212	11,368
Number of Queries per Document	2.05	2.14	2.13
Length of Query (in words)	1.50	1.52	1.52
Length of Summaries (in sentences)	1.26	1.29	1.27

Table 1: Statistics of the synthetic QFS corpus generated from CNN-DailyMail using $k = 3$ (selecting up to 3 queries per document)

News articles are often related to multi-agent real-world events, making them topically diverse documents. Thus, multiple diverse high quality queries can be generated for each document. This is in contrast to other summarization corpora that correspond to topically narrow document classes like WikiHow articles (Koupaee and Wang, 2018; Ladhak et al., 2020). We choose the CNN-DailyMail corpus over other news summarization corpora like XSum (Narayan et al., 2018a), since it contains longer summaries from which multiple query-focused summary subtexts can be extracted.

2.1 Query Focused Corpus

To generate the QFS corpus from the CNN-DailyMail corpus, we perform the following steps:

1. Generate queries from the summary text corresponding to every document in the corpus using a novel query generation framework
2. For each generated query, select the *subset* (potentially of cardinality > 1) of summary

sentences that contain the query to generate the query-focused summary

3. For each document in the corpus, generate QFS triples using the generated queries and their corresponding focused summaries

Our QFS corpus generation framework generates more diverse queries as we consider a broader class of queries than prior work. We also ensure that the generated queries are salient and that the corpus contains summaries of varying length.

2.1.1 Pre-existing Corpora

Hasselqvist et al. (2017) presented a synthetic QFS corpus where queries were named entities in the summary sentences. While named entities are a good class of candidates for queries, they are certainly not representative of all the types of queries one may encounter (for example, "forest fire"). Another drawback of their strategy is that they treat all summary sentences as separate summaries. This entails that (a) the target summaries are short with no diversity in length even though the original CNN-DailyMail corpus contains longer summaries of varying length; and (b) if an entity is present in multiple sentences of the original summary, then multiple targets are created for the same {document, query} pair, each of which is incomplete and sends conflicting signals to the model. Multiple summaries for a single {document, query} pair also means that evaluation is not straightforward as a generated summary could possibly match any of the candidates.

Abdullah and Chali (2020) proposed a query generation strategy where the 5 words from a document's summary that had the highest similarity to the source document were picked as queries. This technique can select non-entities as well and picks candidates that are most relevant to the document as computed by cosine similarity between the query and document. However, the single word restriction means that the generated queries are often fragments of atomic larger queries. For example, names with more than one word ("James Bond") and atomic noun phrases ("dwarf galaxy") are fragmented. Though stop words are removed, there is still the possibility of generating generic low quality queries (like "simply").

2.1.2 Query Generation

We introduce a novel query generation strategy that addresses the limitations of prior techniques, generating queries that are (a) from a broader linguistic class that is more representative of user queries; (b) multi-word phrases; and (c) salient in terms of information content. We base our query generation algorithm on the unsupervised keyphrase extraction technique EmbedRank (Bennani-Smires et al., 2018). EmbedRank generates keyphrases from a single document by extracting candidates, ranking them on document relevance and then removing similar candidates using MMR (Carbonell and Goldstein, 1998) to ensure diversity.

Algorithm 1: Query generation algorithm

```
Input: Text to extract queries from, IDF
       Model, Salient Named Entity Types
Output: List of extracted queries with
        corresponding IDF scores
queries = {};
idf_scores = {};
candidates = Noun-Phrases(Text) ∪
  Named-Entities(Text);
for candidate in candidates do
    Trim leading stopwords in candidate;
    Remove possessive apostrophes in
      candidate;
    Split candidate into contiguous
      sub_spans, where each sub-span is
      either;
        - Salient Named Entity;
        - Proper Noun;
        - Other Remaining;
    Filter sub_spans with more than 5
      words;
    queries ← queries || sub-spans;
end
for query in queries do
    idf_scores ← idf_scores ||
      mean_{word∈query}((idf_word − idf_min)/(idf_max − idf_min));
end
```

Keyphrases generated using EmbedRank suffer from problems like (a) extremely generic keyphrases that should be ignored (e.g., "interesting ones"); (b) stop word prefixes that should be trimmed (e.g.,"other World Cup matches"); and (c) long keyphrases that should be split to avoid highly parochial queries that match with fewer summary sentences during corpus generation (e.g., "energetic new rock band Pearl Jam"). We thus augment this algorithm by making two key modifications. Firstly, we introduce a new algorithm for keyphrase

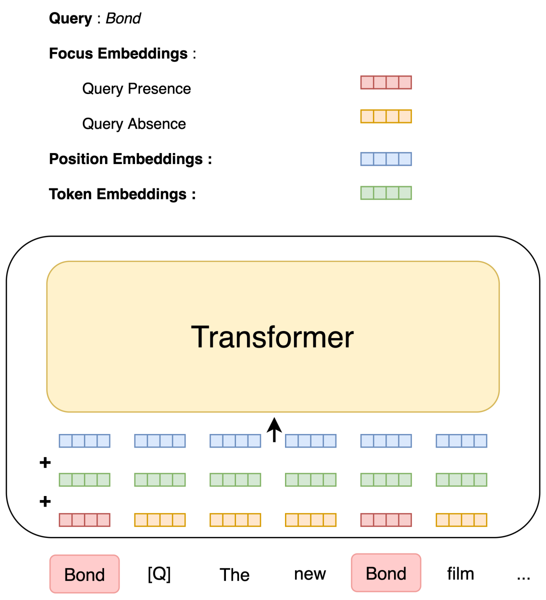

Figure 1: Model architecture with query prefix and focus embeddings

candidate extraction. We use the dependency parse of the document to extract contiguous candidate spans, as shown in Algorithm 1. In addition to this, we generate an aggregate IDF value for each candidate, which is then used to weight the candidate scores while ranking them.

Using this new query generation algorithm, a QFS corpus is created from the CNN-DailyMail dataset. We filter out generated keyphrases with scores below a threshold and choose the top-k remaining keyphrases as queries for each document. Some statistics on this generated corpus using $k = 3$ are shown in Table 1. It is interesting to note that as compared to the base corpus, position information in our synthetic corpus is a weaker signal since the first few sentences in the document may not necessarily be relevant to a query of interest. Query relevance is a stronger signal along with salience and we propose a summarization model that captures this information effectively.

2.2 Cross-Lingual Setting

Since we want to train a summarization model that is capable of handling poor translations during inference time, we follow Ouyang et al. (2019) and perform round trip translation (RTT) to generate noisy English versions of the corpus. The documents in the synthetic QFS corpus are first translated to a foreign pivot language and then back translated to English.

3 Model

We present a novel query-focused summarization model that is built using pre-trained encoder-decoder transformers. Lewis et al. (2020) introduced a denoising pre-training strategy for training sequence-to-sequence models for language generation (BART). Inspired by its success on the generic summarization task, we use BART as our pre-trained transformer model. The pre-trained decoder is useful as the parameter weights learned from the denoising pre-training tasks are a good starting point for fine-tuning the model to produce fluent summaries even when the inputs to the encoder are noisy. We add query focus to BART by introducing two key modifications to the encoder: (i) prefixing the document with the query to contextualize the document embeddings on the query as well; and (ii) adding a new set of embeddings called focus embeddings, in addition to BART's token and position embeddings, to encode the input.

3.1 Prefixing Document with Query

Prefixing the document with the query before encoding it leverages the self-attention mechanism in a transformer to generate document embeddings that are not just contextualized on its content but also on the query. In our model, the query is added to the beginning of the document and delimited by a special separator $[Q]$.

3.2 Focus Embeddings

In addition to query prefixing, we also introduce query focus by explicitly marking the query tokens wherever they appear in the document. We use a new set of embeddings, called *focus embeddings*, which embed query and non-query tokens differently. Introducing new embedding layers has been shown to be effective in providing external knowledge for entity linking and document clustering (Logeswaran et al., 2019; Saravanakumar et al., 2021). For each token in the input, BART uses the summation of two embeddings - token and position - as the input to the first transformer layer. We augment this with an additional embedding layer where tokens in the input text that appear in the query are assigned to one embedding vector while all other tokens are assigned to another vector, as schematically shown in Figure 1. These embeddings are learned during fine-tuning and the model thus learns to project the query terms in the input differently and thereby add focus to those tokens.

	ROUGE 1	ROUGE 2	ROUGE L
Hasselqvist et al.†	13.04	2.29	11.60
BART (Lewis et al., 2020)†	23.62	8.92	20.63
BART with Constrained Decoding (Mao et al., 2020)†	25.60	8.47	20.98
Our Model	37.31	18.79	32.92

Table 2: Automatic summarization metrics on our generated QFS corpus using $k = 1$ (single query per document). † indicates significant difference between baseline and our model (with $p < 0.01$)

	ROUGE 1	ROUGE 2	ROUGE L
Hasselqvist et al.†	13.01	2.66	12.13
Our Model	37.61	19.09	33.12

Table 3: Automatic summarization metrics on our generated QFS corpus using $k = 3$ (up to 3 queries per document). † indicates significant difference between baseline and our model (with $p < 0.001$)

	Relevance	Self-BLEU
Hasselqvist et al.	19.28	30.28
Our Model	96.95	16.06

Table 4: Query focus evaluation metrics on our generated QFS corpus using $k = 3$ (up to 3 queries per document)

	ROUGE 1	ROUGE 2	ROUGE L
Hasselqvist et al.†	6.30	4.14	5.80
Mao et al.†	17.45	4.37	13.97
Our Model	20.50	6.21	17.98

Table 5: Automatic summarization metrics on the cross-lingual DUC2004 dataset. † indicates significant difference between baseline and our model (with $p < 0.01$)

4 Experiments and Results

4.1 Corpus Generation

As mentioned in Section 2.1.2, the generation of the QFS corpus involves selecting the top-k keyphrases with a score above a threshold. In our implementation, we set this threshold to 0.7, which was determined experimentally through a human evaluation of the generated keyphrases. We have two versions of the corpus for $k = 1$ and $k = 3$. We use the *en_core_web_sm* spaCy model (Honnibal et al., 2020) for dependency parsing and named entity detection. [3]

To generate the round trip translated version of the CNN-Daily Mail corpus, we use the open-source Opus MT models (Tiedemann and Thottingal, 2020) for translation with greedy decoding. We use Arabic as the pivot foreign language, consistent with the DUC 2004 (Over and Yen, 2004) Task 3 dataset we use during evaluation.

4.2 Summarization

We use the BART-base as our pre-trained transformer model and randomly initialize a new focus embedding matrix. The query focused model is then trained on our generated CNN-Daily Mail QFS corpus for a maximum of 6 epochs. Early stopping is implemented with validation check done every 10000 steps. Training is done with an effective batch size of 256 using gradient accumulation, learning rate of 5e-4, dropout with $p = 0.1$, label smoothing with $\alpha = 1$, adafactor optimizer and half-precision floating points. Validation checks are done using greedy decoding. The input to the model is trimmed to 512 tokens and the target summaries are trimmed to 128 tokens.

4.3 Results

Baselines The task of cross-lingual QFS, to our best knowledge, hasn't been attempted before. However, we compare our model against prior work on query-focused summarization. In the pre-transformer era, Hasselqvist et al. (2017) introduced a GRU-based pointer generator network architecture for QFS. The model followed the encoder-decoder architecture with attention, encoding queries using a separate RNN. Abdullah and Chali (2020) proposed a QFS technique, where the novelty was the permutation of input sentences based on query relevance. They fine-tuned the BertSum (Liu, 2019) model on their permuted input. However, since the newer model BART Lewis et al. (2020) has shown better performance, we use that as a baseline. Our final baseline is the inference-time constrained text generation framework proposed by Mao et al. (2020), where the constraint in the QFS task is the query for which a focused summary is to be generated.

[3] We only use the following entity classes for keyphrase generation - PERSON, NORP, FAC, ORG, GPE, LOC, PRODUCT, EVENT, WORK OF ART, LAW and LANGUAGE - and exclude classes like ORDINAL

Query	Summary
Lithuania	*Output:* England face Lithuania in their Euro 2016 qualifier on Friday night *Gold:* England host Lithuania in their Euro 2016 qualifier on Friday
Daniel Sturridge	*Output:* Daniel Sturridge has been withdrawn from the squad with a hip injury *Gold:* Daniel Sturridge withdrew from the England squad on Monday night

Table 6: Examples of output and gold summaries for the multiple generated queries from a single article. The article in context discusses the replacement of Daniel Sturridge by Harry Kane in England's Euro 2016 qualifier

	Fluency			Relevance		Coverage		
	Fluent	Partially fluent	Not fluent	Relevant	Not relevant	Complete	Partial	Low/No coverage
Hasselqvist et al.	35.33	26.67	38.00	50.00	50.00	25.33	24.00	50.67
Mao et al.	64.66	28.00	07.33	92.00	08.00	40.67	31.33	28.00
Our Model	69.33	27.33	03.33	94.00	06.00	54.67	26.00	19.33

Table 7: Human evaluation results - accuracy of fluency, relevance and coverage as annotated by human judges

Summarization Evaluation We first evaluated our model against all baselines on our generated round-trip-translated QFS corpus with $k = 1$. The results of this experiment are shown in Table 2. It can be seen that our model substantially outperforms all baselines on the ROUGE metrics (Lin, 2004). The BART model performs better than Hasselqvist et al. (2017) because of the rich knowledge gained during pre-training. Model performance is further improved by providing the query as the inference-time constraint. Our results show that training using our query focusing strategies results in state-of-the-art QFS performance on our corpus.

We also compared our model to Hasselqvist et al. (2017) on the corpus generated with $k = 3$ and the results are shown in Table 3. Since the query is not used during training in BART and Mao et al. (2020), we exclude them from this evaluation since the training data sizes aren't comparable. Our model outperformed the baseline by a significant margin.

Query Focus Evaluation In addition to summarization metrics, we also evaluated the query focusing ability of the models using two metrics - query relevance and diversity. We compute both these metrics on the $k = 3$ corpus. Query relevance is computed as the fraction of summaries that contain (ignoring case) the query that was used to produce it. This metric is computed on the summary for every $\{document, query\}$ pair independently and quantifies how well the summaries capture the query.

Another attribute of the QFS model we evaluate is its ability to produce diverse summaries for different queries on the same document. Since the $k = 3$ corpus has documents with multiple queries, diversity is computed on each document (with >1 query) independently. We use the Self-BLEU metric (Zhu et al., 2018) to measure diversity, where a lower score means greater diversity. For this evaluation, we used a subset of test documents that had more than 1 query and computed Self-BLEU on the set of generated summaries across all queries for each document. It is observed that our model outperforms the baseline on both metrics and the results of the query focus evaluation are shown in Table 4. A few sample summaries generated by our model are shown in Table 6.

Cross-Lingual Evaluation To evaluate our model on real-world translation data, we use the DUC 2004 Task 3 dataset, which consists of human-written English summaries for translated Arabic news articles. Since the corpus is not query focused, we pair each summary with the top query generated using our framework. It is noted here that there is no currently available summarization corpus that is both query-focused and cross-lingual. The results of this evaluation are shown in Table 5. It is observed that our model significantly outperforms the baseline, thus demonstrating its real-world performance gains.

Human Evaluation Since the generated summaries are abstractive, we performed an evaluation where we asked human annotators to evaluate summaries on three dimensions - *fluency* (to evaluate how well the model can produce well-formed summaries even though the inputs are poorly translated), *relevance* (to evaluate how focused to the query the summaries are) and *coverage* (to evaluate the completeness of the generated summaries).

	ROUGE 1	ROUGE 2	ROUGE L
Without Query Prefix and Focus Embeddings†	23.97	7.78	20.36
Only Query Prefix†	36.06	17.80	31.82
Query Prefix and Focus Embeddings	37.61	19.09	33.12

Table 8: Ablation study - automatic summarization metrics on our corpus using $k = 3$ (up to 3 queries per document) to evaluate the impact of focus embeddings. † indicates significant difference between the specified and our proposed model with both query prefix and focus embeddings (with $p < 0.01$)

	ROUGE 1	ROUGE 2	ROUGE L
Hasselqvist et al.	18.03	5.04	16.17
Our Model	39.87	22.84	36.00

Table 9: Automatic summarization metrics on the dataset presented in Hasselqvist et al. (2017)

We sampled 50 instances from the test set of the $k = 1$ corpus for the human evaluation. Given a query and a summary, we asked 3 independent annotators to evaluate the summary on the dimensions mentioned above and the aggregate results are shown in Table 7. It is observed that our model outperforms baselines on every dimension, which correlates well with the automatic metrics presented before. Not only does our model produce relevant summaries, but it is also able to outperform baselines in producing fluent summaries from disfluent documents.

4.4 Ablation Studies

Impact of Focus Embeddings Since the self-attention mechanism in transformers is powerful by itself, we evaluated the impact of the focus embeddings to quantify the gain in performance due to their addition. We conducted an ablation study comparing the performance of the model with and without these embeddings. The results of the experiment are shown in Table 8. It can be seen that while query focusing through self-attention yields a large improvement over query-agnostic vanilla BART, the focus embeddings are useful indeed and produce a significant increase in performance.

Impact of Model and Data Since we presented both a new summarization model as well as a dataset for cross-lingual QFS, we evaluated the impact of each on the final results. For this evaluation, we use a version of our $k = 3$ QFS data without doing round-trip translation to introduce disfluencies, making it comparable to prior work.

To evaluate the impact of the proposed model, we trained and tested our QFS model on the Hasselqvist et al. (2017) dataset and the results are shown in Table 9. It can be seen that our model outperforms the baseline, demonstrating the performance gains due to our QFS architecture. We then evaluated the impact of our data generation framework by comparing (a) a model trained on our dataset and evaluated on the Hasselqvist et al. (2017) test data; (b) a model trained on the Hasselqvist et al. (2017) dataset and evaluated on our test data. In addition to the raw ROUGE scores, we also compute the degradation in performance due to cross-corpus transfer, as compared to a model trained on the corresponding in-corpus train set for each test dataset. The goal of this evaluation was to show that our data generation framework is more robust and can transfer well to the Hasselqvist et al. (2017) dataset even though it is out of distribution, in addition to algorithmically subsuming and augmenting prior generation techniques. The results of this evaluation are shown in Table 10. It can be seen that cross-corpus transfer from our data generation framework results in better summarization performance than from the prior framework. It can also be seen that the performance degradation due to cross-corpus transfer from our framework is much lower than from the baseline, demonstrating the robustness of our data generation methodology.

4.5 Future Work

While our technique produces new state-of-the-art results for cross-lingual QFS, there are still further research challenges that will be the focus of future work. Summarization models in a translate-then-summarize pipeline can fix lexical and syntactic disfluencies introduced by the translation model. However, factual inconsistencies are much harder to handle and were not part of the scope of our work. Our proposed query generation methodology improves upon prior work and generates a wider spectrum of queries. But all the generated queries are still lexically limited to the gold summary and aren't thematic abstract queries (for instance, "wellness" and "sport" for an article that talks about mental fatigue among cricket players). Semantic typing

Training data	Test data	ROUGE 1	ROUGE 2	ROUGE L	Δ ROUGE 1	Δ ROUGE 2	Δ ROUGE L
Hasselqvist et al. Data	Our English Data	24.25	10.05	20.87	-16.29	-13.91	-15.86
Our English Data	Hasselqvist et al. Data	34.79	18.69	30.80	-5.08	-4.15	-5.20

Table 10: Automatic summarization metrics by training and evaluating our proposed model architecture on the specified train and test data. The Δ ROUGE scores quantify the degradation in performance due to cross-corpus transfer, as compared to a model trained on the in-corpus training set for each test dataset

of concepts in the summary and performing query expansion are a few ways of synthesizing an even broader class of queries. Finally, the literature on QFS has only considered queries relevant to the document. This can be extended by generating negative examples and training models to detect and generate summaries only for relevant queries. These are some of the interesting directions of research to pursue.

5 Related Work

The task of cross-lingual query-focused abstractive summarization has, to our best knowledge, never been attempted before. However, the individual dimensions of this task have independently been attempted before. The closest related work in the literature is on cross-language sentence selection, which can be thought of as a form of extractive QFS (Chen et al., 2021). Abstractive summarization is the task of paraphrasing the salient contents of a document with potential verbal innovation (Nallapati et al., 2016; Paulus et al., 2017; Gehrmann et al., 2018; Chen and Bansal, 2018; Fabbri et al., 2019). This is in contrast to extractive summarization, which refers to the selection of salient sentences or phrases from a document (Nallapati et al., 2017; Narayan et al., 2018b; Zhou et al., 2018; Liu et al., 2019; Liu and Lapata, 2019). Contemporary work on abstractive summarization has leveraged transformers to achieve state-of-the-art results (Devlin et al., 2019; Khandelwal et al., 2019; Zhang et al., 2020; Qi et al., 2020; Lewis et al., 2020).

Query-focused summarization has been explored in both the single-document (Nema et al., 2017; Egonmwan et al., 2019; Ishigaki et al., 2020; Laskar et al., 2020; Xie et al., 2020; Zhong et al., 2021) and multi-document setting (Feigenblat et al., 2017; Baumel et al., 2018). The task has been modeled similar to the question answering task, with the query being a question and the summary being similar to a terse answer to the question, sourced from the document. The debatepedia corpus (Nema et al., 2017) is a standard single-document QFS corpus that models the task in this manner, where queries are questions (for example, "Economics: is algae biofuel economically viable?"). This style of queries corpus entails that models trained on QA tasks transfer well to summarization on this corpus (Egonmwan et al., 2019; Su et al., 2021). However, this style is unnatural and is markedly distinct from what a user would enter in a search-and-summarize engine. In this paper, we focused on the QFS task where queries are short phrases. The lack of datasets with this style of queries prompted prior work to develop synthetic corpus generation strategies (Hasselqvist et al., 2017; Abdullah and Chali, 2020; Kulkarni et al., 2020).

Cross-lingual summarization techniques have widely adopted the pipeline strategy - performing translation and summarization as independent cascaded steps (Orăsan and Chiorean, 2008; Wan et al., 2010). Recent work has also attempted to perform joint translation and summarization (Wan et al., 2019; Zhu et al., 2020; Cao et al., 2020; Dou et al., 2020), though it is noted here that these techniques were all applied to high-resource languages, mainly Chinese.

6 Conclusion

In this paper, we presented a zero-resource approach to cross-lingual QFS that involved synthetic corpus generation and a query-focused summarization model. We introduced a novel keyphrase generation algorithm that addressed key issues with prior work like expanding scope to non-entities, handling multi-word phrases and excluding generic uninformative queries. Our data generation framework is more robust than prior techniques both algorithmically and in terms of its ability for cross-corpus transfer. Our summarization model, built on the BART transformer model, introduced query focus by leveraging the self-attention mechanism and introducing focus embeddings that highlight query terms in the document. Our model achieves state-of-the-art results on both our corpus and a prior corpus, with substantial gains over baselines on both automatic metrics and qualitative human evaluation.

References

Deen Mohammad Abdullah and Yllias Chali. 2020. Towards generating query to perform query focused abstractive summarization using pre-trained model. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 80–85, Dublin, Ireland. Association for Computational Linguistics.

Tal Baumel, Raphael Cohen, and Michael Elhadad. 2016. Topic concentration in query focused summarization datasets. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2573–2579. AAAI Press.

Tal Baumel, Matan Eyal, and Michael Elhadad. 2018. Query focused abstractive summarization: Incorporating query relevance, multi-document coverage, and summary length constraints into seq2seq models. *arXiv preprint arXiv:1801.07704*.

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 221–229, Brussels, Belgium. Association for Computational Linguistics.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6220–6231, Online. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Yanda Chen, Chris Kedzie, Suraj Nair, Petra Galuscakova, Rui Zhang, Douglas Oard, and Kathleen McKeown. 2021. Cross-language sentence selection via data augmentation and rationale training. 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 3881–3895, Online. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–686, Melbourne, Australia. Association for Computational Linguistics.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: The bible in 100 languages. *Lang. Resour. Eval.*, 49(2):375–395.

Hoa Trang Dang. 2006. Duc 2005: Evaluation of question-focused summarization systems. In *Proceedings of the Workshop on Task-Focused Summarization and Question Answering*, SumQA '06, page 48–55, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zi-Yi Dou, Sachin Kumar, and Yulia Tsvetkov. 2020. A deep reinforced model for zero-shot cross-lingual summarization with bilingual semantic similarity rewards. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 60–68, Online. Association for Computational Linguistics.

Elozino Egonmwan, Vittorio Castelli, and Md Arafat Sultan. 2019. Cross-task knowledge transfer for query-based text summarization. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 72–77, Hong Kong, China. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Guy Feigenblat, Haggai Roitman, Odellia Boni, and David Konopnicki. 2017. Unsupervised query-focused multi-document summarization using the cross entropy method. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 961–964.

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Johan Hasselqvist, Niklas Helmertz, and Mikael Kågebäck. 2017. Query-based abstractive summarization using neural networks. *arXiv preprint arXiv:1712.06100*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Tatsuya Ishigaki, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, and Manabu Okumura. 2020. Neural query-biased abstractive summarization using copying mechanism. In *Advances in Information Retrieval*, pages 174–181, Cham. Springer International Publishing.

Urvashi Khandelwal, Kevin Clark, Dan Jurafsky, and Lukasz Kaiser. 2019. Sample efficient text summarization using a single pre-trained transformer. *arXiv preprint arXiv:1905.08836*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. Aquamuse: Automatically generating datasets for query-based multi-document summarization. *arXiv preprint arXiv:2010.12694*.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4034–4048, Online. Association for Computational Linguistics.

Md Tahmid Rahman Laskar, Enamul Hoque, and Jimmy Huang. 2020. Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In *Advances in Artificial Intelligence*, pages 342–348, Cham. Springer International Publishing.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *arXiv preprint arXiv:1903.10318*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yang Liu, Ivan Titov, and Mirella Lapata. 2019. Single document summarization as tree induction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1745–1755, Minneapolis, Minnesota. Association for Computational Linguistics.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Yuning Mao, Xiang Ren, Heng Ji, and Jiawei Han. 2020. Constrained abstractive summarization: Preserving factual consistency with constrained generation. *arXiv preprint arXiv:2010.12723*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018a. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018b. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759, New Orleans, Louisiana. Association for Computational Linguistics.

Preksha Nema, Mitesh M. Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1063–1072, Vancouver, Canada. Association for Computational Linguistics.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *Proceedings of* the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2025–2031, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Over and James Yen. 2004. An introduction to duc-2004. *National Institute of Standards and Technology*.

Ramakanth Pasunuru, Asli Celikyilmaz, Michel Galley, Chenyan Xiong, Yizhe Zhang, Mohit Bansal, and Jianfeng Gao. 2021. Data augmentation for abstractive query-focused multi-document summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13666–13674.

Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.

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.

Kailash Karthik Saravanakumar, Miguel Ballesteros, Muthu Kumar Chandrasekaran, and Kathleen McKeown. 2021. Event-driven news stream clustering using entity-aware contextual embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2330–2340, Online. Association for Computational Linguistics.

Dan Su, Tiezheng Yu, and Pascale Fung. 2021. Improve query focused abstractive summarization by incorporating answer relevance. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3124–3131, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926, Uppsala, Sweden. Association for Computational Linguistics.

Xiaojun Wan, Fuli Luo, Xue Sun, Songfang Huang, and Jin-ge Yao. 2019. Cross-language document summarization via extraction and ranking of multiple summaries. *Knowledge and Information Systems*, 58(2):481–499.

Yujia Xie, Tianyi Zhou, Yi Mao, and Weizhu Chen. 2020. Conditional self-attention for query-based summarization. *arXiv preprint arXiv:2002.07338*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–663, Melbourne, Australia. Association for Computational Linguistics.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1309–1321, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *SIGIR*.

Imagine you are in front of a news search engine and are interested in news related to the query below. You type in this query on the search engine which then returns a summary of a recent news article related to your query.

Query:

NASA

Summary:

NASA's first failed mission in 10 years ended when the messenger probe crashed into Mercury.

Is the news summary relevant to the query?

- ○ Yes
- ○ No

How would you rate the fluency of the summary?

- ○ The summary is fluent
- ○ The summary is somewhat fluent with minor errors
- ○ The summary is not fluent with major errors

Does the summary support everything in the statement - "NASA's messenger probe crashed into Mercury, thereby ending mission. This has been NASA's first in failure over 10 years."?

- ○ Yes
- ○ Partially
- ○ No

Figure 2: Questionnaire for human judges to evaluate the model output summaries.

Query	Summary	Relevance Rating
UCI	UCI chief executive Richard Scudamore to be promoted to chief executive	Relevant
Tottenham	west ham sign harry redknapp's contract with west ham	Irrelevant

Table 11: Examples of results from the human evaluation for query relevance rating

Query	Summary	Fluency Rating
Gibraltar	Gibraltar beat Scotland 85-58 in their Euro 2016 qualifier on Sunday	Fluent
Daniel Kirkwood	Daniel Kirkwood, 18, was stabbed outside nightclub in the early hours of this morning	Somewhat fluent
mineral sands mine	she worked with mining mining mining mining mining mining	Not fluent

Table 12: Examples of results from the human evaluation for fluency rating

Generated Summary	Expected Summary	Coverage Rating
The 26-year-old has been linked with a move to Arsenal and Tottenham	The 26-year-old had been linked with moves to Arsenal and Tottenham	Full coverage
Gibraltar beat Scotland 85-58 in their Euro 2016 qualifier on Sunday	Scotland face Gibraltar on Sunday, while Northern Ireland are at home to Finland on the same day	Partial Coverage
The Western Australian now works at a mineral sands mine in Cataby	The 25-year-old mineral sands mine was replaced by Shane Moeman	Low/No coverage

Table 13: Examples of results from the human evaluation for coverage rating

A Human Evaluation

In this section, we provide additional details on the human evaluation conducted on the output summaries from our model and the baselines. The questionnaire given to the human judges is shown in Figure 2. The judges are given a query and the generated summary and asked to rate the summary on fluency, query relevance and coverage. While query relevance was a binary question, fluency and coverage were ternary questions with an in-between option. Examples from the human evaluation results where the human annotators gave different ratings along the three dimensions are shown in Tables 11, 12 and 13.