Constrained Regeneration for Cross-Lingual Query-Focused Extractive Summarization

Elsbeth Turcan[1], **David Wan**[2,*], **Faisal Ladhak**[1], **Petra Galuscakova**[3,†], **Sukanta Sen**[4,‡]
Svetlana Tchistiakova[4,‡], **Weijia Xu**[5], **Marine Carpuat**[5], **Kenneth Heafield**[4],
Douglas Oard[5] and **Kathleen McKeown**[1]
[1]Department of Computer Science, Columbia University;
[2]Department of Computer Science, University of North Carolina at Chapel Hill;
[3]Université Grenoble Alpes; [4]School of Informatics, The University of Edinburgh;
[5]Department of Computer Science & UMIACS, The University of Maryland
{eturcan, faisal}@cs.columbia.edu, davidwan@cs.unc.edu

Abstract

Query-focused summaries of foreign-language, retrieved documents can help a user understand whether a document is actually relevant to the query term. A standard approach to this problem is to first translate the source documents and then perform extractive summarization to find relevant snippets. However, in a cross-lingual setting, the query term does not necessarily appear in the translations of relevant documents. In this work, we show that constrained machine translation and constrained post-editing can improve human relevance judgments by including a query term in a summary when its translation appears in the source document. We also present several strategies for selecting only certain documents for regeneration which yield further improvements.

1 Introduction

Query-focused summarization creates an overview of a document which reflects how that document may be relevant to a provided query; such a task is useful for any search engine, such as for news articles or academic papers, where a user may want to search documents by a given query. In this paper, we further narrow the use case to one in which the user seeks a document containing a single specific query term (which may be a multi-word expression). For example, if the query is "dossier", the user is interested in finding information about a specific type of collection of files, as might exist in an intelligence investigation. A summary in the user's language can help them decide if a foreign language document is relevant.

Our work focuses on query-focused extractive summarization in a cross-lingual setting, where the summaries are generated in a language (here, English) different from the source language of the documents (here, Farsi, Kazakh, and Georgian). Because large summarization corpora do not exist in these languages, we follow a translate-then-summarize approach (Wan et al., 2010) in which we first apply machine translation (MT) to translate documents into English, a language with abundant summarization corpora, and then summarize the translated document; however, this introduces additional concerns. Translating a document once, before a query term is known, can lead to wording choices that are sub-optimal for any particular query term (e.g., if the Kazakh for "dossier" were translated as "file", and it may be unclear whether the specific meaning of dossier occurred in the source as opposed to other meanings of "file"). To address this, we present a *constrained regeneration* framework where we translate a document, summarize it with an extractive summarizer that uses evidence from the source language, and select a sentence to be regenerated under the constraint to include the requested query term if appropriate.

Our work is implemented within a pipeline that includes cross-lingual information retrieval (CLIR) followed by summarization of retrieved documents; in the latter step, a summary is generated for a document given a specific query term. Based on the intuition that seeing the query term in the summary is a strong signal of relevance to end users, we first present work on three types of constrained regeneration systems: Marian-C, a constrained version of Marian (Junczys-Dowmunt et al., 2018); EDITOR (Xu and Carpuat, 2021); and constrained automatic post-editing (Wan et al., 2020, cAPE). In initial experimentation, however, we found that these systems often insert the requested query term even in cases when the foreign document did not contain a suitable translation. To address this, we further in-

*Work performed while at Columbia University
†Work performed while at the University of Maryland
‡Work performed while at the University of Edinburgh

Proceedings of the 29th International Conference on Computational Linguistics, pages 2668–2680
October 12–17, 2022.

troduce document selection methods to determine when to apply regeneration and thus avoid inserting query terms inappropriately. We perform a human evaluation and show that the combined use of regeneration and document selection improve humans' ability to accurately distinguish relevant and irrelevant non-English documents by their generated English summaries.

Our approach combines complementary strengths of the three primary modules needed for cross-lingual query-focused summarization: CLIR excels at discovering cross-lingual mappings at the lexical level, neural MT produces complete sentences that are often very fluent, but sometimes at the expense of adequacy and term preservation, and summarization helps users assess relevance efficiently. The novelty of our approach lies in a tight integration of these components, exploiting CLIR to detect relevance, and combining summarization and selective regeneration of summary sentences to produce a human-useful summary.

Our contributions are as follows:

1. An approach to cross-lingual query-focused summarization using constrained regeneration to make it easier for humans to detect relevant documents.
2. A method of document selection enabling selective application of constrained regeneration to avoid over-generation of the query term.
3. Human evaluation in three different languages demonstrating that constrained MT performs better than constrained automatic post-editing for low-resource settings and that we can improve the end user's ability to identify relevant documents using our approach.

2 Background

2.1 Problem Definition

In this work, we operate in the setting of cross-lingual information retrieval and summarization. Our work focuses primarily on the summarization component of this problem, where we are given an English document D, composed of multiple sentences $s_1, s_2, ..., s_n$, and a search query q (which is a text string) and asked to generate a summary of D that condenses the information relevant to q. We apply extractive summarization, which means that our output summary S will be a subset of the sentences in D. In our setting, the English document D is actually a translation of a document F in another language, and the document-query pair $(\langle D, F\rangle, q)$ has been generated automatically by a CLIR system which was given q and a corpus (of length m) of source-language documents and their English translations $U = \{F_i, D_i\}_{0<i\leq m}$. This introduces some uncertainty as to whether D is always truly relevant to q. Moreover, the retrieval system uses a range of methods to deal with the mismatch between the q, D and F vocabulary, such as embeddings, n-best translations, query translation, and query expansion, and the retrieval thus does not guarantee that translations of the query terms occur in D even for the highly relevant documents.

This setup introduces our two main challenges. First, the initial translation of F into D was done without any query in mind, so it may contain synonyms or paraphrases of q, or it may have been incorrectly translated despite being relevant. Second, the generated summaries cannot always assume $\langle D, F\rangle$ is indeed relevant to q. Our goal is to generate summaries that contain the query q if and only if $\langle D, F\rangle$ is relevant to q without rerunning a large pipeline of CLIR and MT components.

2.2 Cross-Lingual Summarization Pipeline

Our system to translate from non-English documents and English query terms into English summaries is made up of several components developed by participants in the MATERIAL program[1]; its architecture can be seen in Figure 1. Documents are first translated from the source language into English using two different MT systems, Marian (Junczys-Dowmunt et al., 2018); and Google's multilingual neural MT (Google NMT) (Johnson et al., 2017).

The CLIR system, which retrieves relevant documents for a given query term and can work in tandem with MT, consists of a combination of 6 retrieval systems, including (1) statistical ranking (such as language models and BM25 (Robertson et al., 1995)), (2) neural ranking (Chen et al., 2021b), (3) re-ranking of both types, (4) stemming, (5) query expansion (using blind relevance feedback), and (6) document expansion (using DeepCT (Dai and Callan, 2019)). These systems were selected to perform optimally on each language and thus they differ for different languages. CLIR provides the ranking of the documents by relevance to the query, and also the cutoff point above which the documents should be relevant. This cut-

[1] https://www.iarpa.gov/research-programs/material

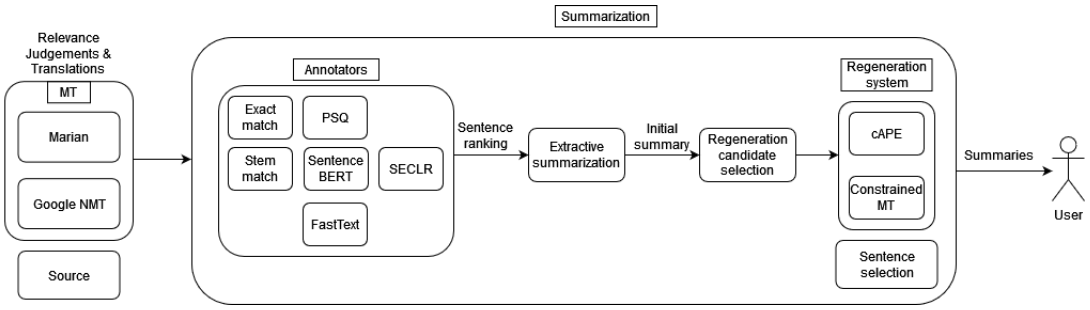

Figure 1: The architecture of our system pipeline as described in subsection 2.2.

System	Marian-C	Marian	EDITOR-C	EDITOR
fa→en	33.1	31.3	26.3	24.8
kk→en	30.2	28.0	20.5	20.5
ka→en	17.6	15.6	25.0	23.4

Table 1: BLEU scores of our constrained and unconstrained MT systems, computed using SacreBLEU (version string BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.5.1)

off uses an average of three estimates – the best ranked cutoff, sum-to-one cutoff and query specific threshold (Zhang et al., 2020) – and it is tuned to achieve an optimal F1 score.

Finally, our summarizer takes a given English document D and English query term q and generates an extractive summary S as relevant as possible to q using sentences from D. The summarizer contains several rankers which each rank all the sentences s_i of D from most to least relevant; these rankers include

1. a count of exact matches to q,
2. a count of stemmed matches to q using MorphAGram (Eskander et al., 2020),
3. mean cosine distance between the translated English sentence s_i and q using the 6B and 42B tokens GloVe embeddings (Pennington et al., 2014),
4. mean cosine distance between the source-language sentence and the English query term q using Probabilistic Structured Queries (PSQ) (Darwish and Oard, 2003),
5. mean cosine distance between the source-language sentence and the translated query term using FastText embeddings (Bojanowski et al., 2016) trained for the source language,
6. cosine distance between the translated English sentence s_i and English query term q using pretrained contextual Sentence-BERT embeddings (Reimers and Gurevych, 2019) based on RoBERTa-large (Liu et al., 2019), and
7. a cross-language sentence selector (Chen et al., 2021a, SECLR) that ranks directly using the sentences in the source language and the query term q.

We combine these rankings using the Borda count algorithm, a standard algorithm for unsupervised combination of rankers (Lillis, 2020; Aslam and Montague, 2001), to obtain a final relevance ranking for all the sentences in the document. We score the output of each MT system separately on each sentence to select the most appropriate translation for a given sentence, and then select the most relevant sentences to add to the summary until a fixed-length word budget is exhausted.

3 Our Models

Our approach to improve relevance judgments of summaries is based on (1) constrained regeneration models that encourage the inclusion of query terms in document translations, and (2) document and sentence selection models that identify documents where the inclusion of query terms is appropriate.

3.1 Constrained Regeneration

We experiment with three approaches that constrain the system to use the query term in the generated summaries: autoregressive MT (section 3.1.1), non-autoregressive MT (section 3.1.2), and automated post-editing (section 3.1.3). These approaches represent diverse state-of-the-art strategies to encourage rather than enforce the inclusion of query terms in translations (i.e., the query terms are soft rather than hard constraints). In this work, we experiment with soft constraints over hard constraints because of the intuition that soft constraints give our models the freedom to choose more natural synonyms and morphology as needed, and and based on empirical evidence that soft constraints result in more fluent and overall better translations (Xu and Carpuat, 2021).

3.1.1 Autoregressive Constrained Machine Translation: Marian-C

Marian-C is a constrained variant of the Marian system (Junczys-Dowmunt et al., 2018) trained on augmented synthetic data to encourage it to include English query terms in the translated English sentence.[2] Following Dinu et al. (2019),[3] we use a data augmentation technique to train our model to copy supplied query terms into its output. Augmentation simply consists in concatenating a query term to the source side of each training sample (with `|||`, a token of three pipe characters, as a delimiter). We create synthetic query terms for our parallel text by extracting random spans of target text of 1 to 3 words. We augment the data in this way with 75% probability. For the remaining 25%, we use the original training sample, to preserve the model's ability to translate when a query term is not available. During inference, the query q is simply appended to the source with the same delimiter.

3.1.2 Nonautoregressive Constrained Machine Translation: EDITOR

EDITOR takes the source sentence $\boldsymbol{x} = (x_1, x_2, ..., x_L)$ (where, in our case, $\boldsymbol{x} = s_i$) and optionally a sequence of constraint terms $\boldsymbol{C} = (c_1, ..., c_m)$ (here, $\boldsymbol{C} = q$) as inputs to generate a translation $\boldsymbol{y}$ that contains most of the constraint terms (Xu and Carpuat, 2021). The output is generated by iteratively editing an input sequence using repositioning, deletion and insertion operations. Constraints are seamlessly incorporated in decoding as the initial sequence $\boldsymbol{y}^0 = \boldsymbol{C}$ to be refined. They can thus be incorporated into the generated translation, or deleted, as the model sees fit. This process does not require custom training. An EDITOR model trained on a standard MT task can incorporate constraints in this way out of the box.

Table 1 shows that the resulting systems provide a wide range of quality levels as measured intrinsically by BLEU (Papineni et al., 2002). Farsi-English is evaluated on IWSLT 2012 and 2013 (Federico et al., 2012; Cettolo et al., 2013), Kazakh-English on WMT 2019 (Barrault et al., 2019), and Georgian-English on the MATERIAL ANALYSIS data described in section 4. Despite using similar parallel training sets, Marian performs better than EDITOR on Farsi (fa) and Kazakh (kk), while EDITOR outperforms Marian on Georgian (ka), reflecting independent system development processes that leverage monolingual data differently. Nevertheless, this provides a wide variety of translations that the summarization model can choose from.

3.1.3 Constrained Automatic Post-Editing

In contrast to MT systems that generate a new translation from scratch, constrained automatic post-editing (**cAPE**) edits the initial translation by incorporating desired words and fixing other potential errors. Following Wan et al. (2020), we use an autoregressive multi-source transformer model for this task. It takes as input the source sentence and the generated translation and outputs the corrected English sentence with the desired query term.

We generate synthetic post-editing triplets for training as follows. We use OPUS and ParaCrawl if available (section 4), resulting in 1.2M training examples for Kazakh-English, and 11.2M for Farsi. Each parallel sentence pair is augmented with an MT output from the relevant MT system (Marian and Google NMT). The original target plays the role of reference even though it was not generated by post-editing. We apply the same terminology set creation strategy and the same set of hyperparameters described in the original paper.

3.2 Document Selection

Due to the difficulty of cross-lingual retrieval and propagation of errors through the pipeline, we are likely to retrieve multiple documents that are not relevant to the given query. This is particularly problematic for regeneration, since the regeneration systems are optimized for including the constraints in their output, and thus may mislead users into judging summaries of irrelevant documents as relevant. Furthermore, regeneration adds additional computational overhead to the system, so we should run it only when we are relatively certain that a source document F is relevant to the query.

Therefore, in order to reduce the number of false positives, we add a document selection step that rescores the relevant documents by integrating scores from the CLIR system as well as the summarization system. In particular, we consider three values: (1) the document score from the SECLR query relevance component (Chen et al., 2021b) of the summarization system, (2) the CLIR system's document score, and (3) a binary variable that indicates whether the CLIR system's document score is above an F1 maximizing cutoff that was tuned on a development set (see Section 4). The new com-

[2] See Appendix A for training details.

[3] Different from Dinu et al. (2019), we do not use additional input factors.

posite score is simply the sum of those three values, all of which are bounded between zero and one.[4] We tune a threshold for the composite score using 100-fold cross-validation to achieve an optimal F1-score on the dev partition.[5] We then develop two systems to make use of this threshold.

+selection: This system presents documents selected for regeneration to human annotators using regenerated summaries and unselected documents using summaries that have not undergone regeneration. We expect that most unselected documents are not relevant, but this system favors high recall of the sort that may be valuable in applications like patent search or intelligence analysis.

+omission: This system assumes all unselected documents are irrelevant, and only asks human annotators for input on the regenerated summaries of documents that were selected. This is because some use cases may prefer higher precision at the cost of lower recall (for example, a casual searcher may prefer not to see irrelevant documents at all).

3.3 Sentence Selection

Once a document has been selected for regeneration, we use PSQ, a component of our CLIR model, to identify the sentences in the summary where it would be most appropriate to insert the query term. We rank each sentence by the maximum PSQ translation probability of any of its words with respect to the query term. We then select the sentence with highest rank (i.e., highest translation probability) to be regenerated; we break ties by the combined ranking of our other rankers as discussed in section 2.2, thus preferring sentences that also appear most conceptually related to the query term. In the event that no translation equivalent can be found through PSQ, regeneration would be aborted and the summary presented as originally created, but this never happens in our dataset.

4 Data

Machine Translation. The training corpora we use for our regeneration MT systems come from the WMT 2019 (Barrault et al., 2019), OPUS (Tiedemann, 2012), and MATERIAL-BUILD[6] parallel datasets for three languages: Farsi (FA), Kazakh (KK), and Georgian (KA). The dataset statistics are given in Table 3. We evaluate our full system on the MATERIAL text dataset consisting of source documents in the specified language as well as collection of English query terms.

Language	Collection	#documents	#queries
Farsi	ANALYSIS	388	221
	DEV	11,662	221
	EVAL	11,640	1264
Kazakh	ANALYSIS	388	400
	DEV	11,622	400
	EVAL	10,815	765
Georgian	ANALYSIS	388	412
	DEV	11,662	412
	EVAL	11,652	842

Table 2: Number of documents and queries for the MATERIAL dataset for evaluation.

	Corpus	#Sentence
	Farsi-English	
Para	OPUS	8.5M
	Hymers	22K
	Mizan	1M
	MATERIAL-BUILD	34K
	ParaCrawl	178K
	Lorelei	59K
	Kazakh-English	
Para	News Commentary	77K
	Wikititles	117K
	Kazakhtv	97K
	Crawl2019	495K
Mono	OPUS	131K
	News2019	20M
	News commentary.v15	608K
	Georgian-English	
Para	OPUS	1.7M
	Crawled	101K
	MATERIAL-BUILD	4K

Table 3: Parallel and monolingual corpora used in training the MT systems. The MATERIAL-BUILD corpus for Kazakh-English is the same as News Commentary.

Cross-Lingual Information Retrieval. The MATERIAL cross-lingual information retrieval dataset is divided into ANALYSIS, DEV, and EVAL, where ANALYSIS is intended for data statistics and examination, DEV for tuning and EVAL for test. The size of the splits are shown for each language in Table 2, and the structure of the data is similar to previous releases (Zavorin et al., 2020). This data includes, for each of our three languages, a separate collection of non-English news and blog documents, a separate collection of English query strings, and gold relevance annotations for each document-query pair within a language.

[4] We tried learning a logistic classifier; however, simply taking the sum of the scores performed similarly.

[5] We take the mean threshold over the 100 folds as the final threshold for our system.

[6] https://www.iarpa.gov/index.php/research-programs/material

5 Experiments

For each of our three languages, we draw a random sample of query-document pairs with a high likelihood of being relevant according to our trained CLIR system for that language. Then, for each query-document pair, we generate an extractive summary with no regeneration applied; this is our **baseline** system. We then apply each applicable constrained regeneration system to each summary independently, generating a new copy of the summary for each regeneration system.

We then submit each of these summaries to Amazon Mechanical Turk for human evaluation, the formulation of which is described in detail below in section 5.1. The result of the human evaluation is a relevance score for each summary–that is, for each $(q, \langle D, F\rangle,$ regeneration system$)$ triple. We can evaluate different regeneration systems at this stage by simply collecting the labels they are assigned and comparing them to the ground truth relevance labels. Finally, we apply our document selection methods; we select a subset of documents whose summaries should be regenerated according to our document selection threshold, and we use the collected scores from the regenerated and baseline variants to evaluate the +selection and +omission variants of the regeneration systems.

Our Farsi experiments are done on a random sample of 1000 query-document pairs from the documents returned by CLIR[7] for the MATERIAL EVAL partition, equally split between ground-truth irrelevant and relevant documents. These query-document pairs are selected such that the summaries the baseline system produced did not contain the query word, indicating an opportunity for regeneration systems to incorporate query terms. We repeat the experiments for Kazakh and Georgian similarly, using samples of 2000 documents for each language from their DEV partitions.

5.1 Human Evaluation

We evaluate our systems in an end-to-end fashion; in our setup, this means that we compare the ground-truth gold relevance label for each query-document pair with the relevance judgment assigned to that pair by human evaluators. The system we develop inherently includes a human in the loop, as its intended purpose is to allow a human to find documents relevant to an intended search term quickly and easily; therefore, we also involve human annotators in its evaluation.

[7]Documents returned by CLIR are those above a threshold that maximizes the Actual Query-Weighted Value (AQWV) that was learned on the dev partition). Details on this metric can be found at https://www.nist.gov/itl/iad/mig/iarpa-material-program.

Score	Precision	Recall	F1
Baseline	53.00	36.81	43.44
Baseline + omission	79.17	13.19	22.62
+cAPE	57.89	76.39	**65.87***
+cAPE +selection	59.79	53.01	56.20*
+cAPE +omission	**81.93**	29.40	43.27*
+Marian-C	52.58	70.83	60.36*
+Marian-C +selection	56.72	52.78	54.68
+Marian-C +omission	72.41	29.17	41.58
+EDITOR	50.22	**79.86**	61.66
+EDITOR +selection	57.89	53.47	55.60
+EDITOR +omission	75.44	29.86	42.79*

Table 4: **Farsi-English Document Relevance Evaluation**. Bold indicates the best score, and stars indicate statistically significant improvement over the baseline (by the approximate randomization test, $p < 0.05$).

For our human evaluation of our summaries, we asked workers on Amazon Mechanical Turk whether generated summaries were relevant to the given query term. We presented the summary, with any exact matches to the query term highlighted in a different color, to workers and asked them to rate the relevance on a five-point scale: {definitely irrelevant, probably irrelevant, unsure, probably relevant, definitely relevant}. For evaluation purposes, each worker's rating was binarized such that "probably relevant" and "definitely relevant" correspond to "relevant", and the others to "irrelevant". We asked three workers to evaluate each summary and aggregated their binarized judgments by majority vote, yielding a single final "relevant" or "irrelevant" human label for each query-summary pair. An example of the interface for this evaluation is included in Appendix B.

5.2 Evaluation Metrics

Our problem is a binary classification problem: a document-query pair is either relevant or irrelevant. We compare relevance judgements obtained during

Score	Precision	Recall	F1
Baseline	25.18	39.08	30.63
Baseline +omission	87.50	16.09	27.18*
+cAPE	25.48	75.86	38.15*
+cAPE +selection	30.61	51.72	38.46*
+cAPE +omission	**89.29**	28.74	43.48*
+Marian-C	21.52	**81.61**	34.05*
+Marian-C +selection	31.37	55.17	40.00
+Marian-C +omission	82.35	32.18	**46.28***
+EDITOR	24.90	68.97	36.59
+EDITOR +selection	26.76	43.68	33.19
+EDITOR +omission	78.26	20.69	32.73*

Table 5: **Kazakh-English Document Relevance Evaluation**. Bold indicates the best score, and stars indicate statistically significant improvement over the baseline (by the approximate randomization test, $p < 0.05$).

Score	Precision	Recall	F1
Baseline	14.35	29.25	19.25
Baseline +omission	30.43	13.21	18.42*
+cAPE	18.11	45.28	25.88
+cAPE +selection	17.02	37.74	23.46
+cAPE +omission	**35.38**	21.70	26.90*
+Marian-C	15.00	**62.26**	24.18*
+Marian-C +selection	18.31	49.06	26.67
+Marian-C +omission	30.70	33.02	**31.82***
+EDITOR	14.44	50.65	22.48
+EDITOR +selection	16.18	42.86	23.49
+EDITOR +omission	31.25	25.97	28.37

Table 6: **Georgian-English Document Relevance Evaluation**. Bold indicates the best score, and stars indicate statistically significant improvement over the baseline (by the approximate randomization test, $p < 0.05$).

human evaluation with reference judgments from the MATERIAL data, using the standard precision, recall and F_1 metrics. Reporting precision and recall independently provides important indicators of the incidence of false positives and false negatives respectively. A false positive represents a document that was not truly relevant to the query, but for which the generated summary falsely convinced the human annotators that it was relevant. Conversely, a false negative represents a relevant document whose summary failed to convey its relevance to the query (and thus human annotators judged it irrelevant). We hypothesize that the blind application of regeneration to even irrelevant documents is likely to decrease the false negative rate, but it may also increase the false positive rate.

We also note that the +selection and +omission systems can be evaluated for each regeneration system by replacing unselected documents' human evaluation with either the human evaluation of the original, non-regenerated document (+selection), or an automatic "irrelevant" judgment (+omission).

6 Results

The results of our experiments are shown in Table 4 (Farsi-English), Table 5 (Kazakh-English), and Table 6 (Georgian-English). Different result trends emerge for the high-resource (Farsi) and low-resource (Kazakh, Georgian) languages.

Beginning with Farsi, we see that applying regeneration via cAPE performs best, improving the F1 score by 20 points over the baseline; both constrained MT systems yield lesser but similar improvements. These improvements are due to dramatic increases in recall and similar precision as compared to the baseline, indicating that relevant documents are much more likely to be noticed and selected by human annotators. In Farsi, however, the additional layer of document selection is unhelpful, as it mitigates the recall too much without a large increase in precision; simply applying regeneration to every returned document-query pair performs best for Farsi.

For the low-resource languages, Kazakh and Georgian, applying regeneration via cAPE or Marian-C shows consistent and significant improvement over the baseline, with Marian-C performing best. EDITOR particularly improves recall over the baseline, but overall the improvements are not statistically significant. We see similar trends as in Farsi, where applying any form of regeneration increases recall and yields similar precision when not using document selection, leading to increased F1. When we include document selection as a pipeline

step before applying regeneration, however, precision also increases for all systems while retaining an improvement in recall (though not as large); as the vast majority of documents are irrelevant to any given query term, selection results in an overall net increase in F1 for cAPE and Marian-C. EDITOR, which is less aggressive in including its constraints, interacts poorly with document selection in Kazakh and yields reduced F1 under this setting. Finally, the +omission variant of document selection actually performs best overall because of how much it improves precision, although it does not increase recall as much as the +selection variant. Thus we see three variants of our systems (no selection, +selection, +omission) occupying different points on the precision-recall tradeoff in the low-resource setting.

We therefore see that for our low-resource languages, adding document selection to our regeneration improves the overall performance because it increases precision; the regeneration systems in these languages tend to take irrelevant documents and make their summaries appear relevant. However, in the case of our high-resource language, the improvements to precision afforded by document selection are minimal and do not balance out its diminished recall. We note that from Table 1, the performance of the base MT systems in Farsi is better than that for the low-resource languages, and correspondingly, the performance of our end-to-end system is best in Farsi, even for the baseline. Our hypothesis is that the documents returned by CLIR for Farsi are already relevant and high-quality compared to those in the low-resource languages; thus document selection helps identify relevant documents in low-resource languages but is not necessary for Farsi.

7 Related Work

Constrained Machine Translation. One of the crucial components of our system is the ability of the MT system to generate translations with specific terminology. Recent works use either constrained decoding, which modifies the decoding scheme to specify which words must be incorporated in the output (Post and Vilar, 2018; Hokamp and Liu, 2017; Hasler et al., 2018), or data augmentation techniques which incorporate the query term as an additional input in the training data (Dinu et al., 2019; Wan et al., 2020; Xu and Carpuat, 2021), avoiding the need to add overhead to the decoding scheme.

Cross-lingual Summarization. Prior work on cross-lingual summarization has mostly focused on two paradigms – summarize-then-translate (Lim et al., 2004; Orăsan and Chiorean, 2008; Wan et al., 2010) and translate-then-summarize (Leuski et al., 2003; Ouyang et al., 2019). The summarize-then-translate approach, however, requires a large amount of summarization training data in the source language (Ladhak et al., 2020), which makes them unsuitable for our setting since the source languages in our setting are low-resource. Prior work has shown that translate-then-summarize approaches are prone to error propagation (Ouyang et al., 2019; Ladhak et al., 2020), and propose methods to produce more fluent summaries. In our setting, having a fluent translation is not sufficient – we also need to have a translation with wording that is appropriate for the given input query. Therefore, in our work we focus on an integration of summarization with regeneration to more clearly indicate relevance.

Query-Focused Summarization. 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; Su et al., 2021) and multi-document setting (Feigenblat et al., 2017; Baumel et al., 2018). Prior work models this task as a 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. Unlike prior work, which has focused on monolingual settings, our work looks at query-focused summarization in the cross-lingual setting, where the query (and therefore the output summary) is in a different language than the source document.

8 Conclusion

We have presented a novel method of cross-lingual query-focused extractive summarization in which we apply regeneration to a generated summary in order to force inclusion of the query term when it appears in the source language document. We demonstrated large, significant improvements over the baseline in all cases through the addition of regeneration, showing increased recall and precision over the baseline. For our noisy low-resource languages, the combination of an aggressive constrained MT system and a document selection filter

additionally allows the benefits of including the query term in a relevant summary while avoiding creating new false positives. We experimented with three methods of constrained regeneration, which attempt to re-translate or edit a given sentence to include a given constraint: constrained automatic post-editing (cAPE), nonautoregressive MT (EDITOR), and our own implementation of autoregressive MT (Marian-C). For low resource languages, autoregressive MT consistently performed better, while for Farsi, cAPE was best. We believe this work opens the door to interesting future work experimenting with more complex varieties of document selection; with different, customized kinds of constrained regeneration; and with what types of languages benefit from these and other techniques.

9 Acknowledgements

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via contract FA8650-17-C-9117. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes not withstanding any copyright annotation therein.

References

Javed A. Aslam and Mark Montague. 2001. Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 276–284, New York, NY, USA. Association for Computing Machinery.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

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*, abs/1801.07704.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT evaluation campaign. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Evaluation Campaign*, Heidelberg, Germany.

Yanda Chen, Chris Kedzie, Suraj Nair, Petra Galuscakova, Rui Zhang, Douglas Oard, and Kathleen McKeown. 2021a. 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.

Yanda Chen, Chris Kedzie, Suraj Nair, Petra Galuscáková, Rui Zhang, Douglas W. Oard, and Kathleen R. McKeown. 2021b. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3881–3895. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2019. Context-aware sentence/passage term importance estimation for first stage retrieval.

Kareem Darwish and Douglas W. Oard. 2003. Probabilistic structured query methods. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, page 338–344, New York, NY, USA. Association for Computing Machinery.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium. 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.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. 2012. Overview of the IWSLT 2012 evaluation campaign. In *Proceedings of the 9th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 12–33, Hong Kong, Table of contents.

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*, SIGIR '17, page 961–964, New York, NY, USA. Association for Computing Machinery.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2018. Non-autoregressive neural machine translation. In *International Conference on Learning Representations*.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems 32*, pages 11181–11191. Curran Associates, Inc.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana. Association for Computational Linguistics.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

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: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, page 174–181, Berlin, Heidelberg. Springer-Verlag.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3th International Conference on Learning Representations*, San Diego, CA, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing*, 2(3):245–269.

David Lillis. 2020. On the evaluation of data fusion for information retrieval. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 54–57, New York, NY, USA. Association for Computing Machinery.

Jung-Min Lim, In-Su Kang, and Jong-Hyeok Lee. 2004. Multi-document summarization using cross-language texts. In *NTCIR*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. 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 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. 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. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

David Wan, Chris Kedzie, Faisal Ladhak, Marine Carpuat, and Kathleen McKeown. 2020. Incorporating terminology constraints in automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1193–1204, Online. Association for Computational Linguistics.

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.

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

Weijia Xu and Marine Carpuat. 2021. EDITOR: An Edit-Based Transformer with Repositioning for Neural Machine Translation with Soft Lexical Constraints. *Transactions of the Association for Computational Linguistics*, 9:311–328.

Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong, and Richard Tong. 2020. Corpora for cross-language information retrieval in six less-resourced languages. In *Proceedings of the workshop on Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)*, pages 7–13, Marseille, France. European Language Resources Association.

Le Zhang, Damianos Karakos, William Hartmann, Manaj Srivastava, Lee Tarlin, David Akodes, Sanjay Krishna Gouda, Numra Bathool, Lingjun Zhao, Richard Schwartz Zhuolin Jiang, and John Makhoul. 2020. The 2019 BBN Cross-lingual Information Retrieval System. In *Proceedings of the Cross-Language Search and Summarization of Text and Speech Workshop*, pages 44–51.

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.

A Constrained Machine Translation Training Frameworks

In the case of Marian-C, we train separate autoregressive unidirectional models for each of Farsi-English, Kazakh-English and Georgian-English. To train our models, we first preprocess the parallel data using Moses (Koehn et al., 2007) punctuation normalization, tokenization, and true-casing. We then create a shared byte-pair encoding vocabulary of 32k tokens following the method of Sennrich et al. (2016), and tokenize our parallel data. We train a Transformer-base model following the method of (Vaswani et al., 2017) using the Marian-NMT framework (Junczys-Dowmunt et al., 2018). The models are trained until BLEU score performance (Papineni et al., 2002; Post, 2018) on the validation set ceases to improve for 15 checkpoints. We use the English→X model to create backtranslations (Edunov et al., 2018) of our monolingual data, and train again on a concatenation of the parallel data and the backtranslations together, in the same way, to create our final X→English models.

In the case of EDITOR, we train separate unidirectional models for Farsi-English, Kazakh-English and Georgian-English using the same preprocessing steps as Marian-C except that we use a shared byte-pair encoding vocabulary of 20k tokens. We apply sequence-level knowledge distillation from autoregressive teacher models as widely used in non-autoregressive generation (Gu et al., 2018, 2019; Xu and Carpuat, 2021). We train a Transformer-base model (Vaswani et al., 2017) using fairseq (Ott et al., 2019). The models are trained using Adam (Kingma and Ba, 2015) with initial learning rate of 0.0005 for maximum 300,000 steps. We select the best checkpoint based on validation BLEU (Papineni et al., 2002).

B Amazon Mechanical Turk Interface

An example of our Amazon Mechanical Turk interface for human evaluation can be seen in Figure 2. Five such questions were presented in each Human Intelligence Task (HIT).

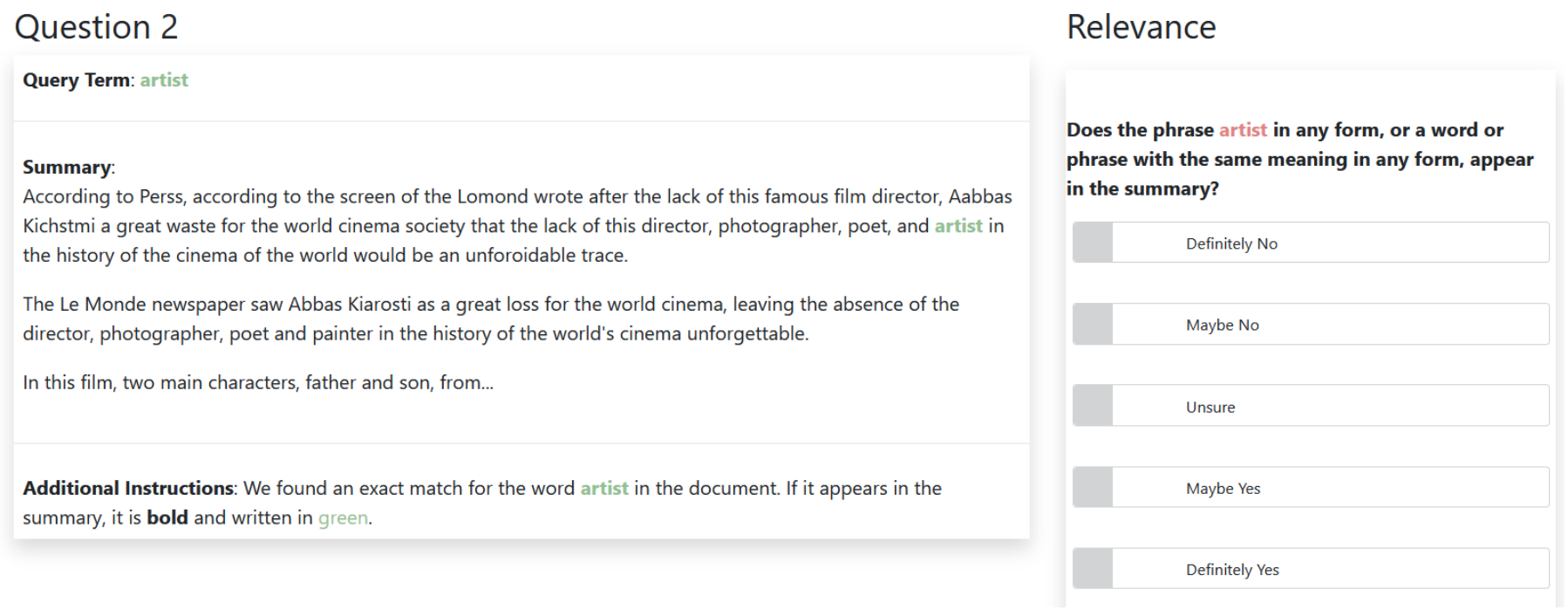

Figure 2: An example interface from our Amazon Mechanical Turk evaluation asking workers whether a given summary is relevant to the query. It includes highlighting of keywords and uses a 5-point scale to evaluate relevance.