A Hybrid Approach to Cross-lingual Product Review Summarization

Saleh Soltan
Alexa AI, New York, USA
ssoltan@amazon.com

Victor Soto
Alexa AI, New York, USA
nvmartin@amazon.com

Ke Tran
Amazon AI Translate, Berlin, Germany
trnke@amazon.de

Wael Hamza
Alexa AI, Dallas, USA
waelhamz@amazon.com

Abstract

We present a hybrid approach for product review summarization which consists of: (i) an unsupervised extractive step to extract the most important sentences out of all the reviews, and (ii) a supervised abstractive step to summarize the extracted sentences into a coherent short summary. This approach allows us to develop an efficient cross-lingual abstractive summarizer that can generate summaries in any language, given the extracted sentences out of thousands of reviews in a source language. In order to train and test the abstractive model, we create the Cross-lingual Amazon Reviews Summarization (CARS) dataset which provides English summaries for training, and English, French, Italian, Arabic, and Hindi summaries for testing based on selected English reviews. We show that the summaries generated by our model are as good as human written summaries in coherence, informativeness, non-redundancy, and fluency.

1 Introduction

Summarizing product reviews with thousands of reviews is a daunting task. At the same time since this task is extremely time consuming to be done by humans, there are no annotated training datasets available for it. Hence, almost all existing approaches rely on unsupervised methods such as Latent Semantic Analysis (LSA) (Steinberger and Jezek, 2004), LexRank (Erkan and Radev, 2004a), MeanSum (Chu and Liu, 2019), and CopyCat (Bražinskas et al., 2020b) to name a few. However, the main two shortcomings of these methods are: (i) they do not provide comparable summaries in terms of coherency and fluency to human written summaries (Bražinskas et al., 2020a), and (ii) they cannot be used for cross-lingual summarization (i.e., provide summaries in a target language given the summaries in the source language).[1]

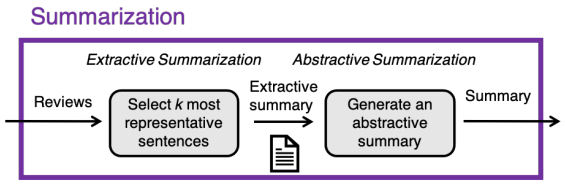

Figure 1: Our proposed summarization pipeline for product review summarization.

In order to address both of these shortcomings, we propose a hybrid two-step summarization approach: (i) an unsupervised extractive summarization step to extract the most representative sentences out all the reviews, and (ii) a supervised abstractive summarization step to summarize the extracted sentences into a coherent short summary.

The main advantage of this approach is that we can rely on a light-weight unsupervised method for the extractive step to reduce the number of reviews and focus on a more expensive supervised model for the abstractive part (and actually collect human summaries for training it). Moreover, we can use state-of-the-art multilingual transformer models to develop a cross-lingual abstractive summarizer relying only on a monolingual extractive step.

For extractive summarization, we use LSA (Steinberger and Jezek, 2004) which is computationally efficient and provides a good performance in our application. To create training data for the abstractive step, we select 1000 products, for each product extract 10 most informative sentences out of all the English reviews using LSA, and obtain 3 summaries per product by asking Amazon Mechanical Turkers to write a short summary of the provided 10 sentences in English (further details are provided in Section 3). An example datapoint is provided in Table A1 in the Appendix.

For testing, we repeat the same process for another 280 products, but this time we ask Turkers to

[1] Although it is possible to machine translate the summaries, it will add an extra inference time and may reduce the quality by translating name of the products etc.

Proceedings of EMNLP 2022 Industry Track, pages 18–28
December 9–11, 2020.

write summaries in French, Spanish, Italian, Arabic, and Hindi (in addition to English) based on the selected English sentences (3 summaries per language per product). We name this dataset Cross-lingual Amazon Reviews Summarization (CARS) dataset.

Finally, we train multiple transformer based models on CARS training data and evaluate their performance on the test data through both automatic and human evaluations. We show that our approach provides informative summaries that are as good as human written summaries and can generalize well to the categories not in the training data (an example of a summary generated by our model is shown in Table A1 in the Appendix).

The main contributions of our work are two fold: 1) introducing a scalable and productionalizable approach for cross-lingual product review summarization capable of producing summaries in English, French, Spanish, Italian, Arabic, and Hindi from English reviews, and 2) demonstrating that our abstractive summarization model outperforms state-of-the-art opinion summarization method (Bražinskas et al., 2020a).

2 Related Work

Some of the most widely used extractive summarization techniques are based on Latent Semantic Analysis (LSA) (Dumais, 2004; Gong and Liu, 2001), in which a matrix that represents the importance of words in sentences is created, and singular value decomposition is used to select the sentences with the highest relevancy; or Bayesian Topic Modeling (Daumé III and Marcu, 2006; Haghighi and Vanderwende, 2009), in which a generative model is used to represent documents as mixtures of latent topics, where a topic is a probability distribution over words. Other extractive methods make use of graph methods (Erkan and Radev, 2004b) and machine learning methods (Wong et al., 2008; Narayan et al., 2018).

Abstractive summarization methods can be structure-based, in which a data structure is used to generate the new summary (examples include tree-based methods (Knight and Marcu, 2000; Kikuchi et al., 2014), template-based methods (Cao et al., 2018) and rule-based methods). More recently, the use of encoder-decoder architectures in transfer learning frameworks have facilitated the use of pre-trained encoders to generate document representations which are then used to generate a new summary (Rush et al., 2015; Chopra et al., 2016; Liu and Lapata, 2019).

Since then there has been increasing efforts to tackle many of the challenges posed by this task: Nallapati et al. (2016) adds linguistically motivated embeddings and pointer networks to deal with out-of-vocabulary words; Paulus et al. (2018) adds reinforcement learning to the training objective to lessen exposure bias; Cohan et al. (2018) proposes a hierarchical model to encode the discourse structure research papers and an attention-based decoder to generate the summaries; Gehrmann et al. (2018) improves content selection performance on neural summarizers by incorporating an extra attention step to constrain on more likely phrases; Desai et al. (2020) incorporates two transformer models to predict the saliency and plausibility of sentence deletion in compressive summarization; Mao et al. (2020) introduces token-level constraints to improve factual consistency. To improve faithfulness Dou et al. (2021) proposes GSUM, a general guided summarization framework that can make use of external guiding policies to ensure faithfulness to the source documents.

On the topic of cross-lingual summarization, Chi et al. (2020) proposes a pre-training strategy for natural language generation tasks on both monolingual and multi-lingual settings followed by monolingual fine-tuning on the downstream task, and show that the resulting model can generalize to new languages. Ouyang et al. (2019) trains low resource cross-lingual summarization systems on automatically translated input and clean references, improving on a standard copy-attention summarizer on low resource languages and also evaluates on an unseen language.

The task of opinion or review summarization, which this paper focuses on, is receiving increased attention due to real-world practical usage. Bražinskas et al. (2020a) proposes FewSUM, a hierarchical framework for few-shot multi-document review summarization that consists on a transformer-based generator followed by plug-in network that switches the generator into a summarizer. FewSUM is the main system we will use for comparison throughout this paper. In a recent concurrent work (Bražinskas et al., 2021), the authors create a dataset of product summaries from a set of product reviews and propose to use joint learning to select a subset of reviews and then summarize from them. However, they use professional written summaries

from various websites as gold summaries (which may not be based on customer reviews at all, leading to model hallucination). The main advantage of our approach in production is its decoupled design. Namely, the the extractive part of our system can always be improved to extract more informative sentences without a need to retrain the abstractive summarizer.

Finally, Gamzu et al. (2021) proposes the task of extreme summarization from multiple product reviews by extracting a single sentence that is concise, relevant and supported by multiple reviews. In our work, we propose to extract the most relevant sentences from each set of reviews, rank them, and use the top 10 reviews to create product summaries. Nevertheless, our decoupled design allows us to use Gamzu et al. (2021)'s method as our extractive summarizer and improve our end to end system in the future.

3 CARS Dataset

In this section, we describe the steps in creating the Cross-lingual Amazon Reviews Summarization (CARS) dataset.

3.1 Products Selection

3.1.1 Train

In order to have a diverse set of products, we selected 1000 products from Electronics, Beauty and Personal Care, Sports, Office Products, and Kitchen categories (200 each) from all of the products with more than 1000 English reviews in the Amazon US marketplace. In each category, we selected 100 products with the average score greater than or equal to 4 out of 5, and 100 products with the average score less than 4 out of 5 (since low-rated products do not have many reviews, we had to use 4 out 5 threshold to separate well-reviewed products from not so well reviewed products).

3.1.2 Test

For test, we selected 280 products from Electronics, Beauty and Personal Care, Sports, Office Products, Kitchen, Apparel, Furniture, Lawn & Garden categories (40 each) with more than 1000 English reviews in the US marketplace (we added 3 extra categories compared to the training set to evaluate generalization ability of models). In each category, we selected 20 products with average score greater than or equal to 4 out of 5, and 20 products with average score less than 4 out of 5.

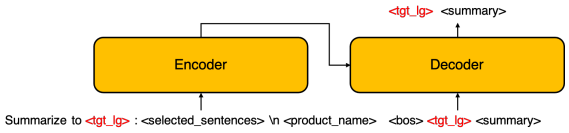

Figure 2: Training MBART50 for abstractive summarization. <tgt_lg> is the target language specific token.

3.2 Extract Sentences

Since products on Amazon have thousands of reviews, it is practically impossible to obtain a gold summary of all reviews for a product. Hence, we extracted a few sentences (out of all reviews) that best describe the important features of a product using an unsupervised extractive summarization method. In particular, we used Latent Semantic Analysis (LSA) (Steinberger and Jezek, 2004) which can run efficiently on reviews of products using a randomized version of the Singular Value Decomposition (SVD) to extract the top 10 sentences representing the reviews for each product (hyperparamter settings are provided in Appendix B). An example of the sentences extracted by LSA is provided in Appendix D.

3.3 Collect Human Summaries

We collected 3 human written summaries for each product in the training data in English and each product in the test data in English, French, Spanish, Italian, Arabic, and Hindi using Amazon Mechanical Turk (AMT). For each product, we provided the 10 selected sentences from all reviews along with the name of the product and asked Turkers to write a short summary of the selected sentences in the target language not exceeding 500 characters. More details on instructions and quality control of the collected summaries is provided in Appendix A.

4 Abstractive Summarizer

4.1 2-Step Training

To train an abstractive summarizer that can generate summaries in any of the target languages given a selected set of sentences in English, we used MBART50 (Liu et al., 2020; Tang et al., 2020) model (which has already been fine-tuned on any-to-any translation task covering 50 languages including the ones presented in the CARS test set) and fine-tuned it for summarization task in 2 steps:

(a) Pre-fine-tuning: We used CNN-DailyMail dataset (Hermann et al., 2015) which includes 290K articles and the corresponding highlights in

English and machine translated all the highlights to Arabic, Hindi, and Italian. Moreover, we used MLSUM dataset (Scialom et al., 2020) which provides 260K article-highlights pair in Spanish and 390K article-highlights pair in French, and machine translated the articles of these datasets from Spanish and French to English. That resulted in a multilingual dataset of English articles to all target languages highlights. We then extracted 10 sentences from each article using LSA to form a dataset of extracted sentences-highlight pairs that mimics the final task. Finally, we fine-tuned the MBART50 on this dataset as shown in Fig. 2.

(b) Fine-tuning: In the second stage of fine-tuning, we fine-tuned the model on CARS training data. Since the training data is in English only, we Machine Translated (MT) the summaries into Spanish, French, Italian, Hindi, and Arabic using Amazon Translate[2] to obtain a cross-lingual dataset from English sentences to summaries in the target languages. We then shuffled the input sentences to obtain two extra versions of the training data (total of 9000 sentences-summary pair for each language). We then fine-tuned the model on the combined dataset of all languages as shown in Fig. 2 with a small addition that we also add the name of the product after the selected sentences as the input to the encoder during fine-tuning. Appendix C provides fine-tuning hyper-parameter details.

4.2 Importance of Including Target Language Token on the Encoder Side

We observed an important property when fine-tuning MBART50. Despite the way MBART50 is fine-tuned for translation by only including the target language token on the decoder side, in our use case, we observed that relying only on the decoder to generate a summary in the target language from a unified representation of the selected sentences (provided by the encoder) significantly degrades the model performance (see Table 1). Hence, we included the target language token on the encoder side as well when we fine-tuned MBART50 for cross-lingual summarization task (as shown in Fig. 2).

5 Evaluation

All the models are trained and evaluated using HuggingFace Transformers Toolkit (Wolf et al., 2020).

[2]https://aws.amazon.com/translate

Model	EN			
	R1	R2	RL	BS
with <tgt_lg> in input	**37.95**	**14.56**	**26.41**	**0.8832**
w/o <tgt_lg> in input	16.34	6.07	11.91	0.8111

Table 1: The importance of including the target language token (<tgt_lg>) in the input (as shown in Fig. 2) when fine-tuning the model on cross-lingual summarization data. **R1** denotes Rouge-1, **R2** denotes Rouge-2, **RL** denotes Rouge-L, and **BS** denotes BERTScore.

5.1 Automatic Evaluation

For evaluation, we included two extra shuffled input sentence orders per gold summary in the test data per language to have a better estimate of the models' performance (9 sentences-summary pairs per product). As in training, we added the name of the product to the end of the selected sentences as the input to the encoder. For evaluation metrics, we use Rouge (Lin, 2004) which is the most common metric for summary evaluation, and BERTScore (Zhang et al., 2020) which has been recently shown to correlates with human judgments the most. For BERTScore, we utilize RoBERTa (Liu et al., 2019) for English and mBERT (Devlin et al., 2019) for non-English summaries (which are the default choices).

Table 2 provides the cross-lingual summarization results on CARS dataset. As can be seen, adding machine translation of English summaries (notice that extracted sentences remain in English) to the training data gives a significant boost to models' performance especially for non-English languages. Moreover, *pre-fine-tuning* the model on a similar task using public data can improve model performance especially in zero-shot case (i.e., training the model only on English summaries). A sample generated summary using the best model (last row in Table 2) is provided in Appendix D.

We also observe that adding machine translated Arabic and Hindi summaries to the training data decreases model performance in Rouge score on these languages (compared to zero-shot with pre-fine-tuning). The main reason for this degradation is that the MT translates the names of the products into Arabic and Hindi (and therefore model learns to translate them as well), whereas gold summaries have the names in English. However, this degradation is not present in BERTScore which relies on token representations instead of their face value.

To see how well the best model (last row in Table 2) generalizes to new categories (ones not in the

Model	EN				FR				ES			
	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS
Fine-tune only on public data												
MBART50	24.28	7.36	18.46	0.8557	17.13	3.11	11.6	0.6506	15.88	2.63	11.76	0.6534
Fine-tune on all English training data												
MBART50	35.97	13.49	25.5	0.8779	8.04	3.84	7.21	0.6731	7.08	2.97	6.61	0.6778
+ pre-fine-tuning	35.92	13.49	25.52	0.876	12.64	5.15	10.35	0.6756	6.81	2.72	6.36	0.6758
Fine-Tune on all English training data plus translation of summaries in all other languages												
MBART50	37.51	14.17	26.21	0.8824	33.1	9.18	20.72	0.7134	34.19	8.64	21.6	0.7191
+ pre-fine-tuning	**37.95**	**14.56**	**26.41**	**0.8832**	**33.4**	**9.21**	**20.73**	**0.7162**	**34.42**	**8.71**	**21.83**	**0.722**

	IT				AR				HI			
	R1	R2	RL	BS	R1	R2	RL	BS	R1	R2	RL	BS
Fine-tune only on public data												
MBART50	17.52	3.22	13.2	0.6433	11.4	4.87	11.27	0.6317	4.81	1.41	4.73	0.609
Fine-tune on all English training data												
MBART50	6.94	2.93	6.26	0.6692	10.23	6.66	9.89	0.6332	6.01	3.37	5.79	0.6086
+ pre-fine-tuning	18.83	5.24	13.93	0.6805	**24.65**	**14.95**	**24.18**	0.6603	**12.12**	**5.4**	**11.89**	0.627
Fine-Tune on all English training data plus translation of summaries in all other languages												
MBART50	29.26	**7.43**	**19.82**	0.7119	20.11	9.07	19.8	**0.7019**	8.8	2.4	8.72	0.6655
+ pre-fine-tuning	**29.31**	7.42	19.7	**0.7137**	18.38	8.45	18.0	0.6995	8.87	2.62	8.83	**0.6682**

Table 2: Cross-lingual summarization results on the CARS set data averaged over 3 gold summaries and 3 different input sentence orders per product (9 samples per product). **R1** denotes Rouge-1, **R2** denotes Rouge-2, **RL** denotes Rouge-L, and **BS** denotes BERTScore.

training data), we computed per category performance of the model as well (see Table 3). Overall, we did not observe any particular degradation in our model's performance on the categories that did not appear in the training data. This indicates that our model generalizes well across domains.

Category	EN			
	R1	R2	RL	BS
Categories that appeared in the training data				
Electronics	33.52	11.54	22.77	0.8757
Beauty	33.06	11.37	23.72	0.8778
Office Product	34.73	12.95	24.13	0.8764
Sports	40.79	17.55	28.69	0.8889
kitchen	41.23	16.92	28.05	0.8848
Categories that *did not* appear in the training data				
Lawn and Garden	38.97	15.01	27.62	0.885
Furniture	41.55	16.06	28.57	0.8895
Apparel	38.88	14	26.39	0.8849

Table 3: Per category performance of our best model.

5.2 Human Evaluation

To evaluate the summaries generated by our best model (last row in Table 2) by humans, we asked Turkers to compare two summaries (the gold summary from test set and the automatic summary by our model) in four different aspects: coherence, informativeness, non-redundancy, and fluency.

Since a high quality evaluation requires constant monitoring of Turkers, we managed to obtain evaluation for only the full English test set and the 2/3rds of the Spanish test set. We evaluated the results using the Best Worst Scaling (BWS) (Kiritchenko and Mohammad, 2016) scores as presented in Table 4. As can be seen, for both English and Spanish the generated summaries are as good as human written summaries. Surprisingly, Turkers found generated summaries to be much better in Fluency in English compared to human written ones. Although the trend reverses in the Spanish summaries which is expected since our model relies only on machine translated training data for Spanish.

Feature	EN		ES	
	Binary	Multi	Binary	Multi
Coherence	-0.0108	**0.0487**	-0.0446	-0.0694
Informativeness	**0.0195**	**0.1234**	-0.038	-0.0992
non-redundancy	**0.1061**	**0.1266**	-0.0777	-0.1636
Fluency	**0.1450**	**0.2792**	-0.0645	-0.1537

Table 4: BWS scores based on human evaluation of the summaries. Binary BWS denotes the scores aggregated in the standard BWS way (-1 if the human summary is better, and +1 if the automatic summary is better) and Multi BWS denotes scores ranging between -3 to +3.

We also looked at the distribution of the scores on both languages (as presented in Figs. 3 and 4). For English, the automatic summaries are always rated to be "much better" more often than the human summaries, whereas for Spanish the opposite is true. In general Spanish summaries, whether automatic or crowdsourced, tend to rate their scores in the "a bit better" area, especially for informativeness, non-redundancy and fluency.

To see how good the generated summaries reflect the overall consensus over a product, we also asked Turkers to give the products a score from 1

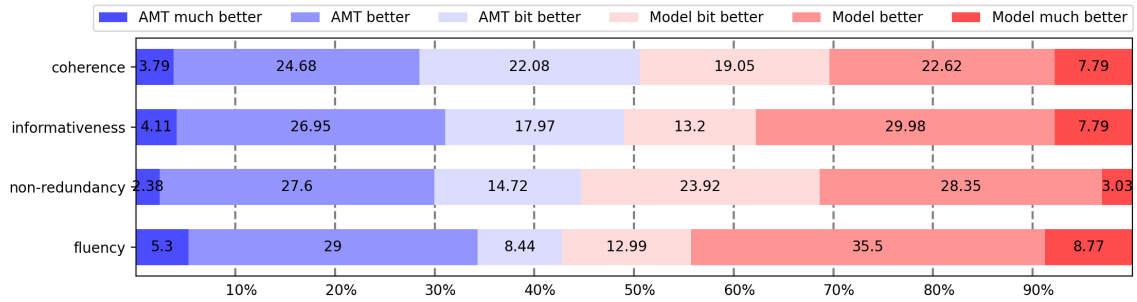

Figure 3: Human comparison between English gold summaries (AMT) and the ones generated by our best model.

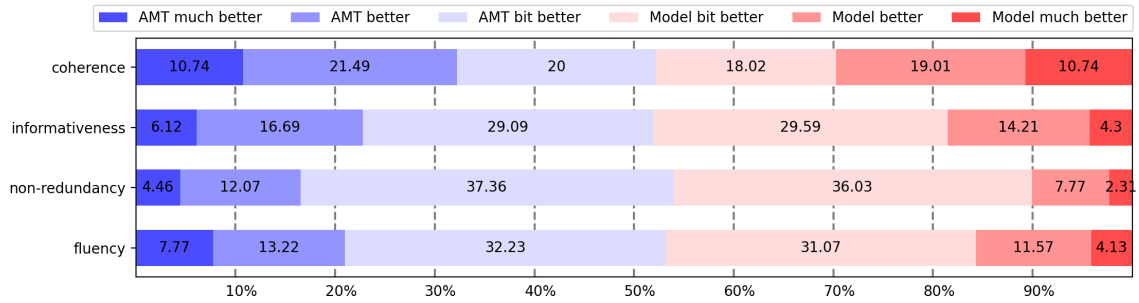

Figure 4: Human comparison between Spanish gold summaries (AMT) and the ones generated by our best model.

Model	R1	R2	RL
FewSum (Bražinskas et al., 2020a)	33.56	7.16	21.49
MBART50	30.12	8.02	19.32
+ pre-fine-tuning	**36.01**	**8.71**	**23.10**

Table 5: Rouge scores on the Amazon reviews test dataset provided in FewSum work. All the models are only fine-tuned on the training data of the same dataset.

to 5 based on the given reviews. We then computed the Mean Absolute Error (MAE) between the real amazon product scores and the guessed scores by the workers from the review summaries. For English, the MAE of both the human and generated summaries are very similar: 0.94 and 0.93, respectively. For Spanish the MAE of the crowdsourced summaries is slightly smaller than the MAE of the generated ones (0.62 compared to 0.75).

5.3 Comparison to FewSum

In the FewSum work (Bražinskas et al., 2020a), authors introduced an unsupervised pretraining strategy specifically for opinion review summarization using hundreds of thousands reviews and demonstrated that their model can provide state-of-the-art summaries only using few supervised review summaries (48 products each having 3 summaries based on 8 randomly selected reviews). However, as can be seen in Table 5, using MBART50 along with our "pre-fine-tuning" strategy (as described in Section 4), our model outperforms FewSum using the exact same training data. Moreover, our model can generate summaries in other languages as well, using only English reviews.

6 Conclusion

We introduced a hybrid approach to cross-lingual product review summarization which provides summaries on different target languages by only relying on English reviews. We demonstrated that our approach results in review summaries that are as good as human written ones in English and Spanish (and comparable to gold summaries in other languages based on automatic evaluation metrics).

We also showed that our pre-fine-tuning plus fine-tuning approach can outperform state-of-the-art in few-shot abstractive review summarization. Moreover, since our abstractive summarizer is trained on summarizing a few selected (maybe unrelated) sentences, our end to end system can be improved by improving the extractive summarization component only without retraining the more expensive multilingual encoder-decoder architecture that we used for abstractive summarization (which is very desirable and cost saving feature in production systems).

The main shortcoming of our work is that it does not provide a mechanism to evaluate the correctness of the generated summaries which is part of our future work.

Acknowledgements

We thank Amjad Jbara and Tobias Falke for their helpful comments and suggestions. We also thank Amjad Jbara, Enrico Piovano, Nicolas Guenon Des Mesnards, and Varun Kumar for their help in data collection and reviewing the summaries written by Turkers in Arabic, Italian, French, and Hindi.

References

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2021. Learning opinion summarizers by selecting informative reviews. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9424–9442, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziqiang Cao, Wenjie Li, Sujian Li, and Furu Wei. 2018. Retrieve, rerank and rewrite: Soft template based neural summarization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 152–161, Melbourne, Australia. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2020. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7570–7577.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Eric Chu and Peter J. Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *Proc. ICML'19*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. 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 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Hal Daumé III and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney, Australia. Association for Computational Linguistics.

Shrey Desai, Jiacheng Xu, and Greg Durrett. 2020. Compressive summarization with plausibility and salience modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6259–6274, Online. 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, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. GSum: A general framework for guided neural abstractive summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4830–4842, Online. Association for Computational Linguistics.

Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Günes Erkan and Dragomir R. Radev. 2004a. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Günes Erkan and Dragomir R. Radev. 2004b. Lexrank: Graph-based lexical centrality as salience in text summarization. 22(1):457–479.

Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying helpful sentences in product reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 678–691, Online. Association for Computational Linguistics.

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.

Yihong Gong and Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, page 19–25, New York, NY, USA. Association for Computing Machinery.

Aria Haghighi and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

K. Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and P. Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. NIPS'15*.

Yuta Kikuchi, Tsutomu Hirao, Hiroya Takamura, Manabu Okumura, and Masaaki Nagata. 2014. Single document summarization based on nested tree structure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 315–320, Baltimore, Maryland. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *arXiv:1412.6980*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. In *Proc. NAACL-HLT'16*.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization-step one: Sentence compression. *AAAI/IAAI*, 2000:703–710.

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.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692*.

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.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

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. 2018. 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.

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.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. *ICLR*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

Josef Steinberger and Karel Jezek. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. ISIM'04*.

Y. Tang, C. Tran, X. Li, P. Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv:2008.00401*.

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

Kam-Fai Wong, Mingli Wu, and Wenjie Li. 2008. Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 985–992, Manchester, UK. Coling 2008 Organizing Committee.

Tianyi Zhang, V. Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *Proc. ICLR'20*.

Appendix

A Data Collection Details

A.1 Instructions

We asked Turkers to follow the following instructions when writing a summary for each product:

1. Summaries written in any language other than the asked language will be rejected.
2. Summaries that do not simplify very long product names will be REJECTED. For example, instead of referring to a hypothetical product as "FakeBrand FakeCode KN95 Protective Mask" you could use "FakeBrand KN95 Mask."
3. Summaries written in first person will be REJECTED. Please write the summary on third person, and never in the first person.
4. Summaries that include information related to shipping or delivery will be REJECTED. Please do not include information related to the shipping or delivery of the product.
5. Summaries that include information related to other products will be REJECTED. Sometimes reviewers might compare the product of interest to other options in the market. It is ok to include how the product of interest fares compared to other options but do not mention specific alternatives.
6. Some opinions might be long, it is ok to synthesize the most important information contained within them.
7. Some opinions might not contain relevant information for a product summary, it is ok to ignore them. For example: "After watching the show, I was looking forward to knowing more about the world, and I decided to give the book a chance."

A.2 Quality Control

Throughout the process, we asked friends and colleagues to approve the quality of the summaries in the target languages (in their native language) before adding them to our train/test set. For all the languages, we could find enough Turkers to write summaries directly in the target languages given the selected sentences in English (although collecting Arabic was extremely slow). However, for Hindi since we were not getting quality summaries in a timely manner, we decided to collect summaries for two thirds of the test data through human translation of the English summaries instead of a combined summarization and translation process that we asked Turkers to do for other languages.

B LSA Hyper-parameters

For LSA, we considered the top 100 unigrams and bigrams appeared in less than 5% of sentences (to avoid most common terms) in the term-frequency matrix and used the top 5 eigenvalues for sorting the most informative sentences (i.e., focused on the 5 most important "topics" for each product). We then used the top 10 sentences as the sentences representing the reviews for each product. Table A1 shows an example of the top sentence in the reviews for a selected product.

C Fine-tuning Details

All the models are trained using Adam (Kingma and Ba, 2015) optimizer with $1e^{-5}$ learning rate, no warm-up steps, a linear learning rate scheduler, and an effective batch size of 112. For pre-fine-tuning, we trained the model for 2 epochs. But for fine-tuning, we trained the models for 5 epochs. For fine-tuning on FewSum data, we used 10 epochs (but as in the case of fine-tuning on CARS data, we included two shuffled versions of the input reviews during training).

D Sample Generated Summaries

A sample data point and generated summaries (not cherry picked) from the test set are provided in Table A1.

Product Name	Loud Alarm Clock with Bed Shaker, Vibrating Alarm Clock for Heavy Sleepers, Deaf and Hard of Hearing, Dual Alarm Clock, 2 Charger Ports, 7-Inch Display, Full Range Dimmer and Battery Backup - Green
Selected Sentences from Reviews	"Outwardly they look great, the large green numbers give excellent readability and thanks to the adjustable brightness of the display, they Shine comfortably at night." The goods have been received Faster than I thought The workmanship is very fine Real time monitoring Uitrasonic alarm Very sensitive High precision A very satisfying shopping "I take sedatives and have slept through friends banging on my bedroom door/windows, fire trucks in my apartment parking lot right outside my door, and massive storms." "First off, I can sleep through a full blown tornado and over the years I've tried every trick in the book to wake me up in the mornings." "My sleep is quite heavy if I work hard during the day, I usually do not even hear my phone alarm or those loud mechanically ringing clocks." "Easy-to-see numbers with adjustable control, alarm sound, light weight, slim, occupying little space, it charges my mobile phone." "I put it on the bedside table, it is easy to read, USB plug to keep your mobile phone, notebook computer and any other you need to recharge." "Overall, I love this product and it's pretty good to wake up sleepy heads in the bed making it a great gift to buy another for my friend." "The extra large display allows even someone like me (-9 power eyesight) can see the time without glasses on, which is a very pleasant thing to experience." I decided on this clock because it had multiple alarm options and it had usb ports to charge my phone at bedside without taking up more plug space.
Human Summary 1	The alarm clock has a loud ringer that will wake you up easily in the morning. The large green numbers on the clock are easy to read and it is easy to adjust the brightness of the display. The clock is small and takes up little space and there is even a USB port to charge devices.
Human Summary 2	The Loud Alarm Clock is designed to wake up heavy sleepers. It vibrates. It has two ports for charging, a seven-inch display, and a battery back up. Reviewers were pleased with the lightweight and slim design, the two chargers, the loud alarm sound and the large, easy-to-read display. There were no negative reviews.
Human Summary 3	This alarm clock is generally well liked by users. They mention that they love the connection ports and it does a good job waking them up.
Generated Summary English	The Loud Alarm Clock with Bed Shaker is a good product, easy to see with adjustable control, the numbers give excellent readability and the brightness of the display is adjustable. The product is lightweight and slim, occupying little space and it is easy to read. It has multiple alarm options and it has USB ports to charge your phone without taking up more plug space.
Generated Summary Spanish	El reloj de alarma ruidoso con sofocantedor de cama es una gran adición a su hogar. Es ideal para personas que tienen problemas de audición graves o no pueden oír el sonido de la alarma cuando dormen. Es fácil de leer y tiene una pantalla de 7 pulgadas que se puede utilizar para cargar sus teléfonos y portátiles. Los números verdes son fáciles de ver con un control ajustable y son ligeros y ligeros, lo que lo convierte en un gran regalo.
Generated Summary French	L'horloge d'alarme forte avec shaker de lit est un excellent ajout à votre équipement de nuit. Il est facile à lire et dispose d'un grand écran. Les chiffres verts sont faciles à lire. Ils s'allument confortablement la nuit et sont livrés avec une grande luminosité réglable. Il y a également deux ports USB pour charger votre téléphone portable et votre ordinateur portable.
Generated Summary Italian	L'orologio svegliare forte con Shaker Bed è un bel prodotto, facile da vedere con un controllo regolabile e un suono di allarme. Il prodotto è leggero e sottile, occupando poco spazio ed è facile da leggere. Il display è extra grande e permette anche a qualcuno di vedere l'ora senza occhiali accesi.

Table A1: An example of test data with the generated summaries in latin languages.