

Few-Shot Learning for Opinion Summarization

Arthur Bražinskas¹ Mirella Lapata¹ Ivan Titov^{1,2}

¹ILCC, University of Edinburgh

²ILLC, University of Amsterdam

abrazinskas@ed.ac.uk, {mlap, ititov}@inf.ed.ac.uk

Abstract

Opinion summarization is the automatic creation of text reflecting subjective information expressed in multiple documents, such as user reviews of a product. The task is practically important and has attracted a lot of attention. However, due to the high cost of summary production, datasets large enough for training supervised models are lacking. Instead, the task has been traditionally approached with extractive methods that learn to select text fragments in an unsupervised or weakly-supervised way. Recently, it has been shown that abstractive summaries, potentially more fluent and better at reflecting conflicting information, can also be produced in an unsupervised fashion. However, these models, not being exposed to actual summaries, fail to capture their essential properties. In this work, we show that even a handful of summaries is sufficient to bootstrap generation of the summary text with all expected properties, such as writing style, informativeness, fluency, and sentiment preservation. We start by training a conditional Transformer language model to generate a new product review given other available reviews of the product. The model is also conditioned on review properties that are directly related to summaries; the properties are derived from reviews with no manual effort. In the second stage, we fine-tune a plug-in module that learns to predict property values on a handful of summaries. This lets us switch the generator to the summarization mode. We show on Amazon and Yelp datasets that our approach substantially outperforms previous extractive and abstractive methods in automatic and human evaluation.

1 Introduction

Summarization of user opinions expressed in online resources, such as blogs, reviews, social media, or internet forums, has drawn much attention due to its potential for various information access applications, such as creating digests, search, and report

Gold	These shoes run true to size , do a good job supporting the arch of the foot and are well-suited for exercise . They're good looking, comfortable , and the sole feels soft and cushioned. Overall they are a nice, light-weight pair of shoes and come in a variety of stylish colors.
Ours	These running shoes are great! They fit true to size and are very comfortable to run around in . They are light weight and have great support . They run a little on the narrow side , so make sure to order a half size larger than normal .
Reviews	perfect fit for me ... supply the support that I need ... are flexible and comfortable It is very comfortable ... I enjoy wearing them running running shoes ... felt great right out of the box ... They run true to size my feet and feel like a dream ... Totally light weight shoes run small ... fit more true to size ... fit is great! ... supports my arch very well They are lightweight... usually wear a size women's 10 ... ordered a 10.5 and the fit is great!

Table 1: Example summaries produced by our system and an annotator; colors encode its alignment to the input reviews. The reviews are truncated, and delimited with the symbol ‘||’.

generation (Hu and Liu, 2004; Medhat et al., 2014; Angelidis and Lapata, 2018).

Although significant progress has been observed in supervised summarization in non-subjective single-document context, such as news articles (Rush et al., 2015; Nallapati et al., 2016; Paulus et al., 2017; See et al., 2017; Liu et al., 2018), modern deep learning methods rely on large amounts of annotated data that are not readily available in the opinion-summarization domain and expensive to produce. A key obstacle making data annotation expensive is that annotators need to consider multiple input texts when writing a summary, which is

time-consuming. Moreover, annotation would have to be undertaken for multiple domains as online reviews are inherently multi-domain (Blitzer et al., 2007) and summarization systems can be domain-sensitive (Isonuma et al., 2017). This suggests that it is unlikely that human-annotated corpora large enough for training deep models will be available.

Recently, a number of unsupervised abstractive multi-document models were introduced (e.g., COPYCAT; Bražiński et al. 2020 and MEANSUM; Chu and Liu 2019) that are trained on large collections of unannotated product reviews.¹ However, unsurprisingly perhaps, since the models are not exposed to the actual summaries, they are unable to learn their key characteristics. For instance, MEANSUM (Chu and Liu, 2019) is prone to producing summaries that contain a significant amount of information that is unsupported by reviews; COPYCAT generates summaries that are better aligned with reviews, yet they are limited in detail. Moreover, both systems, are trained mostly on subjectively written reviews, and as a result, tend to generate summaries in the same writing style.

The main challenge in the absence of large annotated corpora lies in successful utilization of scarce annotated resources. Unlike recent approaches to language model adaptation for abstractive single-document summarization (Hoang et al., 2019; Raffel et al., 2019) that utilize hundreds of thousands of summaries, our two annotated datasets consist of only 60 and 100 annotated data-points. It was also observed that a naive fine-tuning of multi-million parameter models on small corpora leads to rapid over-fitting and poor generalization (Vinyals et al., 2016; Finn et al., 2017). In this light, we propose a few-shot learning framework and demonstrate that even a tiny number of annotated instances is sufficient to bootstrap generation of the formal summary text that is both informative and fluent (see Table 1). To the best of our knowledge, this work is the first few-shot learning approach applied to summarization.

In our work, we observe that reviews in a large unannotated collection vary a lot; for example, they differ in style, the level of detail, or how much they diverge from other reviews of the product in terms of content and overall sentiment. We refer to individual review characteristics and their relations to other reviews as *properties* (Ficler and Goldberg,

2017). While reviews span a large range of property values, only a subset of them is appropriate for summaries. For example, summaries should be close to the product’s reviews in content, avoid using the first-person pronouns and agree with the reviews in sentiment. Our approach starts with estimating a property-aware model on a large collection of reviews and then adapts the model using a few annotator-created summaries, effectively switching the generator to the summarization regime. As we demonstrate in our experiments, the summaries do not even have to come from the same domain.

More formally, we estimate a text model on a dataset of reviews; the generator is a Transformer conditional language model (CLM) that is trained with a ‘leave-one-out’ objective (Besag, 1975; Bražiński et al., 2020) by attending to other reviews of the product. We define properties of unannotated data that are directly related to the end task of summarization. Those properties are easy to derive from reviews, and no extra annotation effort is required. The CLM is conditioned on these properties in training. The properties encode partial information about the target review that is being predicted. We capitalize on that by fine-tuning parts of the model jointly with a tiny *plug-in network* on a handful of human-written summaries. The plug-in network is trained to output property values that make the summaries likely under the trained CLM. The plug-in has less than half a percent of the original model’s parameters, and thus is less prone to over-fitting on small datasets. Nevertheless, it can successfully learn to control dynamics of a large CLM by providing property values that force generation of summaries. We shall refer to the model produced using the procedure as **Few Shot Summarizer** (FEWSUM).

We evaluate our model against both extractive and abstractive methods on Amazon and Yelp human-created summaries. Summaries generated by our model are substantially better than those produced by competing methods, as measured by automatic and human evaluation metrics on both datasets. Finally, we show that it allows for successful cross-domain adaption. Our contributions can be summarized as follows:

- we introduce the first few-shot learning framework for abstractive opinion summarization;
- we demonstrate that the approach substantially outperforms extractive and abstractive

¹For simplicity, we use the term ‘product’ to refer to both Amazon products and Yelp businesses.

models, both when measured with automatic metrics and in human evaluation;

- we release datasets with abstractive summaries for Amazon products and Yelp businesses.²

2 Unsupervised Training

User reviews about an entity (e.g., a product) are naturally inter-dependent. For example, knowing that most reviews are negative about a product’s battery life, it becomes more likely that the next review will also be negative about it. To model inter-dependencies, yet to avoid intractabilities associated with undirected graphical models (Koller and Friedman, 2009), we use the leave-one-out setting (Besag, 1975; Bražinskas et al., 2020).

Specifically, we assume access to a large corpus of user text reviews, which are arranged as M groups $\{r_{1:N}\}_{j=1}^M$, where $r_{1:N}$ are reviews about a particular product that are arranged as a *target review* r_i and $N - 1$ *source reviews* $r_{-i} = \{r_1, \dots, r_{i-1}, r_{i+1}, \dots, r_N\}$. Our goal is to estimate the conditional distribution $r_i|r_{-i}$ by optimizing the parameters θ as shown in Eq. 1.

$$\begin{aligned}\theta^* &= \arg \max_{\theta} \frac{1}{M N} \sum_{j=1}^M \sum_{i=1}^N \log p_{\theta}(r_i^j | r_{-i}^j) \\ &= \arg \max_{\theta} \frac{1}{M N} \sum_{j=1}^M \sum_{i=1}^N \log G_{\theta}(r_i^j | E_{\theta}(r_{-i}^j))\end{aligned}\quad (1)$$

Our model has an encoder-generator Transformer architecture (Vaswani et al., 2017), where the encoder E_{θ} produces contextual representations of r_{-i} that are attended by the generator G_{θ} , which in-turn is a conditional language model predicting the target review r_i , estimated using teacher-forcing (Williams and Zipser, 1989). An illustration is presented in Fig. 1.

The objective lets the model exploit common information across reviews, such as rare brand names or aspect mentions. For example, in Fig. 1, the generator can directly attend to the word *vacuum* in the source reviews to increase its prediction probability.

Additionally, we condition on partial information about the target review r_i using an oracle

$q(r_i, r_{-i})$ as shown in Eq. 2.

$$\frac{1}{M N} \sum_{j=1}^M \sum_{i=1}^N \log G_{\theta}(r_i^j | E_{\theta}(r_{-i}^j), q(r_i^j, r_{-i}^j)) \quad (2)$$

We refer to this partial information as *properties* (Ficler and Goldberg, 2017), which correspond to text characteristics of r_i or relations between r_i and r_{-i} . For example, one such property can be the ROUGE score (Lin, 2004) between r_i and r_{-i} , which indicates the degree of overlap between r_i and r_{-i} . In Fig. 1, a high ROUGE value can signal to the generator to attend the word *vacuum* in the source reviews instead of predicting it based on language statistics. Intuitively, while the model observes a wide distribution of ROUGE scores during training on reviews, during summarization in test time we can achieve a high degree of input-output text overlap by setting the property to a high value. We considered three types of properties.

Content Coverage: ROUGE-1, ROUGE-2, and ROUGE-L between r_i and r_{-i} signals to G_{θ} how much to rely on syntactic information in r_{-i} during prediction of r_i . *Writing Style*: as a proxy for formal and informal writing style, we compute pronoun counts, and create a distribution over 3 points of view and an additional class for cases with no pronouns; see Appendix 9.7 for details. *Rating and Length Deviations*: for the former, we compute the difference between r_i ’s rating and the average r_{-i} rating; in the latter case, we use the difference between r_i ’s length and the average r_{-i} length.

2.1 Novelty Reduction

While summary and review generation are technically similar, there is an important difference that needs to be addressed. Reviews are often very diverse, so when a review is predicted, the generator often needs to predict content that is not present in source reviews. On the other hand, when a summary is predicted, its semantic content always matches the content of source reviews. To address this discrepancy, in addition to using the ROUGE scores, as was explained previously, we introduce a *novelty reduction* technique, which is similar to label smoothing (Pereyra et al., 2017).

Specifically, we add a regularization term \mathcal{L} , scaled by λ , that is applied to word distributions

²Both the code and datasets are available at: <https://github.com/abrazinskas/FewSum>

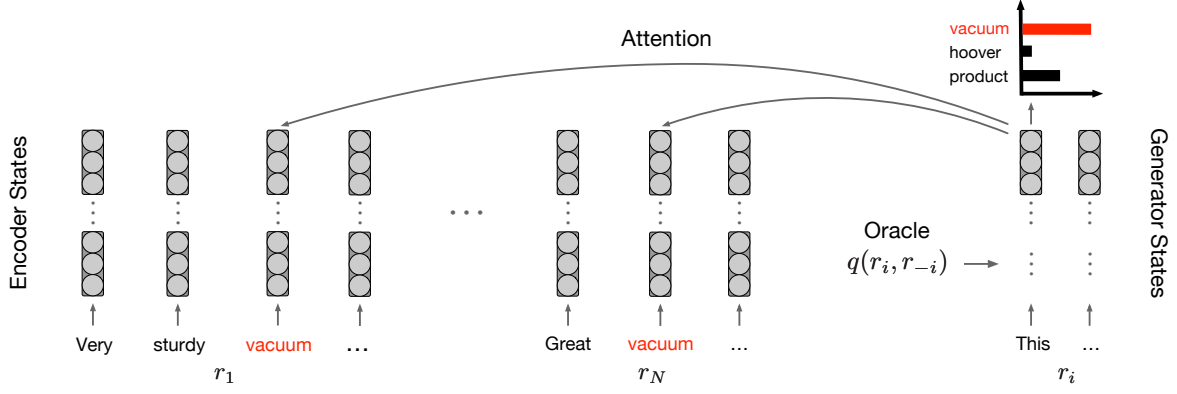


Figure 1: Illustration of the FEWSUM model that uses the leave-one-out objective. Here predictions of the target review r_i is performed by conditioning on the encoded source reviews r_{-i} . The generator attends the last encoder layer’s output to extract common information (in red). Additionally, the generator has partial information about r_i passed by the oracle $q(r_i, r_{-i})$.

produced by the generator G_θ as shown in Eq. 3.

$$\frac{1}{M N} \sum_{j=1}^M \sum_{i=1}^N \left[\log G_\theta(r_i^j | E_\theta(r_{-i}^j), q(r_i^j, r_{-i}^j)) - \lambda \mathcal{L}(G_\theta(r_i^j | E_\theta(r_{-i}^j), q(r_i^j, r_{-i}^j))) \right] \quad (3)$$

It penalizes assigning the probability mass to words not appearing in r_{-i} , as shown in Eq. 4, and thus steers towards generation of text that is more grounded in content of r_{-i} .

$$\mathcal{L}(G_\theta(r_i | r_{-i}, q(r_i, r_{-i}))) = \sum_{t=1}^T \sum_{w \notin V(r_{-i})} G_\theta(W_t = w | r_i^{1:t-1}, r_{-i}, q(r_i, r_{-i})) \quad (4)$$

Here, T is the length of r_i , and the inner sum is over all words that do not appear in the word vocabulary of r_{-i} . Intuitively, in Fig. 1, the penalty could reduce the probability of the word *hoover* to be predicted as it does not appear in the source reviews.

3 Summary Adaptation

Once the unsupervised model is trained on reviews, our task is to adapt it to generation of summaries. Here, we assume access to a small number of annotator-written summaries $(s^k, r_{1:N}^k)_{k=1}^K$ where s is a summary for $r_{1:N}$ input reviews. As we will show in Sec. 6.1, naive fine-tuning of the unsupervised model on a handful of annotated data-points leads to poor generalization. Instead, we capitalize on the fact that the generator G_θ has observed a wide range of property values associated with r_i during the unsupervised training phase. Intu-

itively, some combinations of property values drive it into generation of text that has qualities of a summary while others of a review. However, we might not know values in advance that are necessary for generation of summaries. Furthermore, $q(r_i, r_{-i})$ cannot be applied at test time as it requires access to target texts. In the following section, we describe a solution that switches the generator to the summarization mode relying only on input reviews.

3.1 Plug-in Network

We start by introducing a parametrized *plug-in network* $p_\phi(r_{-i})$ that yields the same types of properties as $q(r_i, r_{-i})$. From a practical perspective, the plug-in should be input-permutation invariant and allow for an arbitrary number of input reviews (Zaheer et al., 2017). Importantly, the trainable plug-in can have a marginal fraction of the main model’s parameters, which makes it less prone to over-fitting when trained on small datasets. We initialize the parameters of $p_\phi(r_{-i})$ by matching its output to $q(r_i, r_{-i})$ on the unannotated reviews. Specifically, we used a weighted combination of distances as shown for one group of reviews in Eq. 5.

$$\sum_{i=1}^N \sum_{l=1}^L \alpha^l D^l(p_\phi(r_{-i})^l, q(r_i, r_{-i})^l) \quad (5)$$

Here, $D^l(p_\phi(r_{-i})^l, q(r_i, r_{-i})^l)$ is a distance for the property l , and α^l is an associated weight. Specifically, we used L1 norm for *Content Coverage*, *Rating and Length Deviations*, and Kullback-Leibler divergence for *Writing Style*.

For the plug-in network, we employed a multi-

layer feed-forward network with multi-head attention modules over the encoded states of the source reviews at each layer, followed by a linear transformation, predicting property values. Note that the encoder is shared with the main model.

3.2 Fine-Tuning

Unsurprisingly, perhaps, the network p_ϕ being initialized on unannotated reviews inherits a strong bias towards outputting property values resulting in generation of reviews, which should not be appropriate for generating summaries. Fortunately, due to the simplicity of the chosen properties, it is possible to fine-tune p_ϕ to match the output of q on the annotated data $(s^k, r_{1:N}^k)_{k=1}^K$ using Eq. 5.

An alternative is to optimize the plug-in to directly increase the likelihood of summaries under G_θ while keeping all other parameters fixed.³

As the generator is trained on unannotated reviews, it might not encounter a sufficient amount of text that is written as a summary, and that highly overlaps in content with the input reviews. We address that by unfreezing the attention module of G_θ over input reviews and the plug-in p_ϕ , and by maximizing the likelihood of summaries:

$$\frac{1}{K} \sum_{k=1}^K \left[\log G_\theta(s^k | E_\theta(r_{1:N}^k), p_\phi(r_{1:N}^k)) \right] \quad (6)$$

This allows the system to learn an interaction between G_θ and p_ϕ . For example, what property values are better associated with summaries and how G_θ should better respond to them.

4 Experimental Setup

4.1 Dataset

For training we used customer reviews from Amazon (He and McAuley, 2016) and Yelp.⁴ From the Amazon reviews we selected 4 categories: *Electronics*; *Clothing*, *Shoes and Jewelry*; *Home and Kitchen*; *Health and Personal Care*. We used a similar pre-processing schema as in (Bražinskas et al., 2020), details are presented in Appendix 9.1. For training, we partitioned business/product reviews to the groups of 9 reviews by sampling without replacement. Thus, for unsupervised training in Sec. 2, we conditioned on 8 reviews for each target review. The data-statistics are shown in Table 2.

³We explored that option, and observed that it works similarly, yet leads to a slightly worse result.

⁴<https://www.yelp.com/dataset/challenge>

Dataset	Training	Validation
Yelp	38,913/1,016,347	4,324/113,886
Amazon	182,932/3,889,782	9,629/205,992

Table 2: Data statistics after pre-processing. The format in the cells is Businesses/Reviews and Products/Reviews for Yelp and Amazon, respectively.

We obtained 480 human-written summaries (180 for Amazon and 300 for Yelp) for 8 reviews each, using Amazon Mechanical Turk (AMT). Each product/business received 3 summaries, and averaged ROUGE scores are reported in the following sections. Also, we reserved approximately $\frac{1}{3}$ for testing and the rest for training and validation. The details are in Appendix 9.2.

4.2 Experimental Details

For the main model, we used the Transformer architecture (Vaswani et al., 2017) with trainable length embeddings and shared parameters between the encoder and generator (Raffel et al., 2019). Subwords were obtained with BPE (Sennrich et al., 2016) using 32000 merges. Subword embeddings were shared across the model as a form of regularization (Press and Wolf, 2017). For a fair comparison, we approximately matched the number of parameters to COPYCAT (Bražinskas et al., 2020). We randomly initialized all parameters with Glorot (Glorot and Bengio, 2010). For the plug-in network, we employed a multi-layer feed-forward network with multi-head attention modules over encoded states of the source review. After the last layer, we performed a linear projection to compute property values. Further, parameter optimization was performed using Adam (Kingma and Ba, 2014), and beam search with n-gram blocking (Paulus et al., 2017) was applied to our model and Copycat for summary generation.

All experiments were conducted on 4 x GeForce RTX 2080 Ti.

4.3 Hyperparameters

Our parameter-shared encoder-generator model used a 8-head and 6-layer Transformer stack. Dropout in sub-layers and subword embeddings dropout was both set to 0.1, and we used 1000 dimensional position-wise feed-forward neural networks. We set subword and length embeddings to 390 and 10 respectively, and both were concatenated to be used as input. For the plug-in network, we set the output dimension to 30 and internal feed-forward network hidden dimensions to 20. We used

a stack of 3 layers, and the attention modules with 3 heads at each layer. We applied 0.4 internal dropout and 0.15 attention dropout. Property values produced by the plug-in or oracle were concatenated with subword and length embeddings and linearly projected before being passed to the generator. In total, our model had approximately 25M parameters, while the plug-in network only 100K (i.e., less than 0.5 % of the main model’s parameters).

In all experiments, the hyperparameter tuning was performed based on the ROUGE-L score on Yelp and Amazon validation sets.

4.4 Baselines

LEXRANK (Erkan and Radev, 2004) is an unsupervised extractive graph-based algorithm selecting sentences based on graph centrality. Sentences represent nodes in a graph whose edges have weights denoting similarity computed with tf-idf.

MEANSUM is an unsupervised abstractive summarization model (Chu and Liu, 2019) that treats a summary as a discrete latent state of an autoencoder. The model is trained in a multi-task fashion with two objectives, one for prediction of reviews and the other one for summary-reviews alignment in the semantic space using the cosine similarity.

COPYCAT is the state-of-the-art unsupervised abstractive summarizer (Bražinskas et al., 2020) that uses continuous latent representations to model review groups and individual review semantics. It has an implicit mechanism for novelty reduction and uses a copy mechanism.

As is common in the summarization literature, we also employed a number of simple summarization baselines. First, the CLUSTROID review was computed for each group of reviews as follows. We took each review from a group and computed ROUGE-L with respect to all other reviews. The review with the highest ROUGE score was selected as the clustroid review. Second, we sampled a RANDOM review from each group to be used as the summary. Third, we constructed the summary by selecting the *leading sentences* (LEAD) from each review of a group.

5 Evaluation Results

Automatic Evaluation We report ROUGE F1 score (Lin, 2004) based evaluation results on the Amazon and Yelp test sets in Tables 3 and 4, respectively. The results indicate that our model outperforms abstractive and extractive methods on both

	R1	R2	RL
FewSum	0.3356	0.0716	0.2149
Copycat	0.2785	0.0477	0.1886
MeanSum	0.2663	0.0489	0.1711
LexRank	0.2772	0.0506	0.1704
Clustroid	0.2716	0.0361	0.1677
Lead	0.2700	0.0492	0.1495
Random	0.2500	0.0382	0.1572

Table 3: ROUGE scores on the Amazon test set.

	R1	R2	RL
FewSum	0.3729	0.0992	0.2276
Copycat	0.2812	0.0589	0.1832
MeanSum	0.2750	0.0354	0.1609
LexRank	0.2696	0.0493	0.1613
Clustroid	0.2890	0.0490	0.1800
Lead	0.2620	0.0457	0.1432
Random	0.2148	0.0259	0.1387

Table 4: ROUGE scores on the Yelp test set.

datasets. Also, the results are supported by qualitative improvements over other models, see examples in the Appendix.

Best-Worst Scaling We performed human evaluation with the Best-Worst scaling (Louviere and Woodworth, 1991; Louviere et al., 2015; Kiritchenko and Mohammad, 2016) on the Amazon and Yelp test sets using the AMT platform. We assigned multiple workers to each tuple containing summaries from COPYCAT, our model, LEXRANK, and human annotators. The judgment criteria were the following: *Fluency*, *Coherence*, *Non-redundancy*, *Informativeness*, *Sentiment*. Details are provided in Appendix 9.6.

For every criterion, a system’s score is computed as the percentage of times it was selected as best, minus the percentage of times it was selected as worst (Orme, 2009). The scores range from -1 (unanimously worst) to +1 (unanimously best).

The results are presented in Tables 5 and 6 for Amazon and Yelp, respectively. On the Amazon data, they indicate that our model is preferred across the board over the baselines. COPYCAT is preferred over LEXRANK in terms of fluency and non-redundancy, yet it shows worse results in terms of informativeness and overall sentiment preservation. In the same vein, on Yelp in Table 6 our model outperforms the other models.

All pairwise differences between our model and other models are statistically significant at

	Fluency	Coherence	Non-Redundancy	Informativeness	Sentiment
FewSum	0.1000	0.1429	0.1250	0.2000	0.3061
Copycat	-0.1765	-0.5333	-0.2727	-0.7455	-0.7143
LexRank	-0.4848	-0.5161	-0.5862	-0.3488	-0.0909
Gold	0.5667	0.6364	0.6066	0.6944	0.4138

Table 5: Human evaluation results in terms of the Best-Worst scaling on the Amazon test set.

	Fluency	Coherence	Non-Redundancy	Informativeness	Sentiment
FewSum	0.1636	0.1429	0.0000	0.3793	0.3725
Copycat	-0.2000	-0.0769	0.1053	-0.4386	-0.2857
LexRank	-0.5588	-0.5312	-0.6393	-0.6552	-0.4769
Gold	0.5278	0.3784	0.4795	0.6119	0.4118

Table 6: Human evaluation results in terms of the Best-Worst scaling on the Yelp test set.

	Full (%)	Partial (%)	No (%)
FewSum	43.09	34.14	22.76
Copycat	46.15	27.18	26.67

Table 7: Content support on the Amazon test set.

$p < 0.05$, using post-hoc HD Tukey tests. The only exception is non-redundancy on Yelp when comparing our model and COPYCAT (where our model shows a slightly lower score).

Content Support As was observed by Falke et al. (2019); Tay et al. (2019); Bražinskas et al. (2020), the ROUGE metric can be insensitive to hallucinating facts and entities. We also investigated how well generated text is supported by input reviews. We split summaries generated by our model and COPYCAT into sentences. Then for each summary sentence, we hired 3 AMT workers to judge how well content of the sentence is supported by the reviews. Three following options were available. *Full support*: all the content is reflected in the reviews; *Partial support*: only some content is reflected in the reviews; *No support*: content is not reflected in the reviews.

The results are presented in Table 7. Despite not using the copy mechanism, that is beneficial for fact preservation (Falke et al., 2019) and generation of more diverse and detailed summaries (see Appendix), we score on par with COPYCAT.

6 Analysis

6.1 Alternative Adaptation Strategies

We further explored alternative utilization approaches of annotated data-points, based on the same split of the Amazon summaries as explained

in Sec. 4.1. First, we trained a model in an unsupervised learning setting (USL) on the Amazon reviews with the leave-one-out objective in Eq. 1. In this setting, the model has neither exposure to summaries nor the properties, as the oracle $q(r_i, r_{-i})$ is not used. Further, we considered two alternative settings how the pre-trained unsupervised model can be adapted on the gold summaries. In the first setting, the model is fine-tuned by predicting summaries conditioned on input reviews (USL+F). In the second one, similar to Hoang et al. (2019), we performed adaptation in a multi-tasking learning (MTL) fashion. Here, USL is further trained on a mixture of unannotated corpus review and gold summary batches with a trainable embedding indicating the task.⁵ The results are presented in Table 8.

First, we observed that USL generates summaries that get the worst ROUGE scores. Additionally, the generated text tends to be informal and substantially shorter than an average summary, we shall discuss that in Sec. 6.2. Second, when the model is fine-tuned on the gold summaries (USL+F), it noticeably improves the results, yet they are substantially worse than of our proposed few-shot approach. It can be explained by strong influence of the unannotated data stored in millions of parameters that requires more annotated data-points to overrule. Finally, we observed that MTL fails to decouple the tasks, indicated by only a slight improvement over USL.

⁵We observed that the 1:1 review-summary proportion works the best.

	R1	R2	RL
FewSum	0.3356	0.0716	0.2149
MTL	0.2403	0.0435	0.1627
USL+F	0.2823	0.0624	0.1964
USL	0.2145	0.0315	0.1523
Random	0.2500	0.0382	0.1572

Table 8: ROUGE scores on the Amazon test set for alternative summary adaptation strategies.

	1st	2nd	3rd	NoPr	Len
Gold	0.0	1.7	60.0	38.3	0.0
FewSum	0.5	1.3	83.2	15.0	3.4
USL+F	29.7	0.0	45.3	25.0	-28.6
USL	56.7	0.0	43.3	0.0	-32.7
Reviews	49.0	7.3	35.6	8.1	-17.6

Table 9: Text characteristics of generated summaries by different models on the Amazon test set.

6.2 Influence of Unannotated Data

We further analyzed how plain fine-tuning on summaries differs from our approach in terms of capturing summary characteristics. For comparison, we used USL and USL+F, which are presented in Sec. 6.1. Additionally, we analyzed unannotated reviews from the Amazon training set. Specifically, we focused on text formality and the average word count difference (*Len*) from the gold summaries in the Amazon test set. As a proxy for the former, we computed the marginal distribution over points of view (POV), based on pronoun counts; an additional class (*NoPr*) was allocated to cases of no pronouns. The results are presented in Table 9.

First, we observed that the training reviews are largely informal (49.0% and 7.3% for 1st and 2nd POV, respectively). Unsurprisingly, the model trained only on the reviews (USL) transfers a similar trait to the summaries that it generates.⁶ On the contrary, the gold summaries are largely formal - indicated by a complete absence of the 1st and a marginal amount of 2nd POV pronouns. Also, an average review is substantially shorter than an average gold summary, and consequently, the generated summaries by USL are also shorter. Example summaries are presented in Table 10.

Further, we investigated how well USL+F, adapts to the summary characteristics by being ac-

⁶As beam search, attempting to find the most likely candidate sequence, was utilized, opposed to a random sequence sampling, we observed that generated sequences had no cases of the 2nd POV pronouns and complete absence of pronouns (*NoPr*).

Gold	These shoes run true to size, do a good job supporting the arch of the foot and are well-suited for exercise. They're good looking, comfortable, and the sole feels soft and cushioned. Overall they are a nice, light-weight pair of shoes and come in a variety of stylish colors.
FewSum	These running shoes are great! They fit true to size and are very comfortable to run around in. They are light weight and have great support. They run a little on the narrow side, so make sure to order a half size larger than normal.
USL+F	This is my second pair of Reebok running shoes and they are the best running shoes I have ever owned. They are lightweight, comfortable, and provide great support for my feet.
USL	This is my second pair of Reebok running shoes and I love them. They are the most comfortable shoes I have ever worn.

Table 10: Example summaries produced by models with different adaptation approaches.

Domain	In-domain	Cross-domain
Cloth	0.2188	0.2220
Electronics	0.2146	0.2136
Health	0.2121	0.1909
Home	0.2139	0.2250
Avg	0.2149	0.2129

Table 11: In and cross domain experiments on the Amazon dataset, ROUGE-L scores are reported.

tually fine-tuned on them. Indeed, we observed that USL+F starts to shift in the direction of the summaries by reducing the pronouns of the 1st POV and increasing the average summary length. Nevertheless, the gap is still wide, which would probably require more data to be bridged. Finally, we observed that our approach adapts much better to the desired characteristics by producing well-formed summary text that is also very close in length to the gold summaries.

6.3 Cross-Domain

We hypothesized that on a small dataset, the model primarily learns course-grained features, such as common writing phrases, and their correlations between input reviews and summaries. Also, that they, in principle, could be learned from remotely related domains. We investigated that by fine-tuning the model on summaries that are not in the target domain of the Amazon dataset. Specifically,

we matched data-point count for 3/4 domains of training and validation sets to the in-domain Amazon data experiment presented in Sec 5; the test set remained the same for each domain as in the in-domain experiment. Then, we fine-tuned the same model 5 times with different seeds per target domain. For comparison, we used the in-domain model which was used in Amazon experiments in Sec. 5. We computed the average ROUGE-L score per target domain, where overall σ was 0.0137. The results are reported in Table 11.

The results indicate that the models perform on-par on most of the domains, supporting the hypothesis. On the other hand, the in-domain model shows substantially better results on the *health* domain, which is expected, as, intuitively, this domain is the most different from the rest.

7 Related Work

Extractive weakly-supervised opinion summarization has been an active area of research. LEXRANK (Erkan and Radev, 2004) is an unsupervised extractive model. OPINOSIS (Ganesan et al., 2010) does not use any supervision and relies on POS tags and redundancies to generate short opinions. However, this approach is not well suited for the generation of coherent long summaries and, although it can recombine fragments of input text, it cannot generate novel words and phrases. Other earlier approaches (Gerani et al., 2014; Di Fabrizio et al., 2014) relied on text planners and templates, which restrict the output text. A more recent extractive method of Angelidis and Lapata (2018) frames the problem as a pipeline of steps with different models for each step. Isonuma et al. (2019) introduce an unsupervised approach for single product review summarization, where they rely on latent discourse trees. The most related unsupervised approach to this work is our own work, COPYCAT (Bražinskas et al., 2020). Unlike that work, we rely on a powerful generator to learn conditional spaces of text without hierarchical latent variables. Finally, in contrast to MEANSUM (Chu and Liu, 2019), our model relies on inductive biases without explicitly modeling of summaries. A concurrent model DENOISESUM (Amplayo and Lapata, 2020) uses a syntactically generated dataset of source reviews to train a generator to denoise and distill common information. Another parallel work, OPINIONDIGEST (Suhara et al., 2020), considers controllable opinion aggregation and is a

pipeline framework for abstractive summary generation. Our conditioning on text properties approach is similar to Fierler and Goldberg (2017), yet we rely on automatically derived properties that associate a target to source, and learn a separate module to generate their combinations. Moreover, their method has not been studied in the context of summarization.

8 Conclusions

In this work, we introduce the first to our knowledge few-shot framework for abstractive opinion summarization. We show that it can efficiently utilize even a handful of annotated reviews-summary pairs to train models that generate fluent, informative, and overall sentiment reflecting summaries. We propose to exploit summary related properties in unannotated reviews that are used for unsupervised training of a generator. Then we train a tiny plug-in network that learns to switch the generator to the summarization regime. We demonstrate that our approach substantially outperforms competitive ones, both abstractive and extractive, in human and automatic evaluation. Finally, we show that it also allows for successful cross-domain adaptation.

Acknowledgments

We would like to thank members of Edinburgh NLP group for discussion, and the anonymous reviewers for their valuable feedback. We gratefully acknowledge the support of the European Research Council (Titov: ERC StG BroadSem 678254; Lapata: ERC CoG TransModal 681760) and the Dutch National Science Foundation (NWO VIDI 639.022.518).

References

- Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. *Proceedings of Association for Computational Linguistics (ACL)*.
- Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686.
- Julian Besag. 1975. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.
- Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of Association for Computational Linguistics (ACL)*.
- Eric Chu and Peter Liu. 2019. Meansum: a neural model for unsupervised multi-document abstractive summarization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 1223–1232.
- Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12.
- Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. pages 54–63.
- Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479.
- Tobias Falke, Leonardo FR Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220.
- Jessica Fidler and Yoav Goldberg. 2017. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark. Association for Computational Linguistics.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1126–1135. JMLR. org.
- Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A graph based approach to abstractive summarization of highly redundant opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 340–348.
- Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T Ng, and Bitaj Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1602–1613.
- Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.
- Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.
- Andrew Hoang, Antoine Bosselut, Asli Celikyilmaz, and Yejin Choi. 2019. Efficient adaptation of pre-trained transformers for abstractive summarization. *arXiv preprint arXiv:1906.00138*.
- Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.
- Masaru Isonuma, Toru Fujino, Junichiro Mori, Yutaka Matsuo, and Ichiro Sakata. 2017. Extractive summarization using multi-task learning with document classification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2101–2110.
- Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2019. Unsupervised neural single-document summarization of reviews via learning latent discourse structure and its ranking. In *Proceedings of Association for Computational Linguistics (ACL)*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Svetlana Kiritchenko and Saif M Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 811–817.

- Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries acl. In *Proceedings of Workshop on Text Summarization Branches Out Post Conference Workshop of ACL*, pages 2017–05.
- Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. In *Proceedings of International Conference on Learning Representations (ICLR)*.
- Jordan J Louviere, Terry N Flynn, and Anthony Alfred John Marley. 2015. *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.
- Jordan J Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. *University of Alberta: Working Paper*.
- Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, 5(4):1093–1113.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, 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.
- Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and hb. *Sequim, WA: Sawtooth Software*.
- Romain Paulus, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*.
- Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.
- Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 157–163.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.
- 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.
- Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of Association for Computational Linguistics (ACL)*.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Proceedings of Association for Computational Linguistics (ACL)*.
- Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. Opiniondigest: A simple framework for opinion summarization. *Proceedings of Association for Computational Linguistics (ACL)*.
- Wenyi Tay, Aditya Joshi, Xiuzhen Jenny Zhang, Sarvnaz Karimi, and Stephen Wan. 2019. Red-faced rouge: Examining the suitability of rouge for opinion summary evaluation. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*, pages 52–60.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638.
- Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.
- Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. 2017. Deep sets. In *Advances in neural information processing systems*, pages 3391–3401.

9 Appendices

9.1 Dataset Pre-Processing

We selected only Amazon products and Yelp businesses with minimum of 10 reviews, and the minimum and maximum lengths of 20 and 70 words, respectively. Also, popular products/businesses above the 90th percentile were removed. From each business/product we sampled 9 reviews without replacement to form groups of reviews.

9.2 Evaluation Data Split

From the Amazon annotated dataset, We used 28, 12, 20 products for training, validation, and testing, respectively. On Yelp, we used 30, 30, 40 for training, validation, and testing, respectively. Both the automatic and human evaluation experiments were performed on the test sets.

9.3 Training Procedure

First, to speed-up the training phase, we trained an unconditional language model for 13 epoch on the Amazon reviews with the learning rate (LR) set to $5 * 10^{-4}$. On Yelp we trained it for 27 epochs with LR set to $7 * 10^{-4}$. The language model was used to initialize both the encoder and generator of the main model.

Subsequently, we trained the model using Eq. 2 for 9 epochs on the Amazon reviews with $6 * 10^{-5}$ LR, and for 57 epochs with LR set to $5 * 10^{-5}$. Additionally, we reduced novelty using Eq. 4 by training the model further for 1 epoch with 10^{-5} LR and $\lambda = 2$ on Amazon; On Yelp we trained for 4 epochs, with $3 * 10^{-5}$ LR, and $\lambda = 2.5$.

For the plugin network’s initialization, as explained in Sec. 3.1, we performed optimization by output matching with the oracle for 11 epochs on the unannotated Amazon reviews with $1 * 10^{-5}$ LR. On Yelp we trained for 87 epochs with $1 * 10^{-5}$. Lastly, we fine-tuned the plugin network on the human-written summaries by output matching with the oracle⁷. On the Amazon data for 98 epochs with $7 * 10^{-4}$, and for 62 epochs with $7 * 10^{-5}$ on Yelp. We set weights to 0.1, 1., 0.08, 0.5 for length deviation, rating deviation, POV, and ROUGE scores, respectively. Then fine-tuned the attention part of the model and the plug-in network jointly for 33 epochs with $1 * 10^{-4}$ on the Amazon data. And 23 epochs with $1 * 10^{-4}$ LR on Yelp.

⁷We set rating deviation to 0 as summaries do not have associated human-annotated ratings.

9.4 Summary Annotation

For summary annotation, we reused 60 Amazon products from Bražinskas et al. (2020) and sampled 100 businesses from Yelp. We assigned 3 Mechanical Turk workers to each product/business, and instructed them to read the reviews and produce a summary text. We used the following instructions:

- The summary should reflect user common opinions expressed in the reviews. Try to preserve the common sentiment of the opinions and their details (e.g. what exactly the users like or dislike). For example, if most reviews are negative about the sound quality, then also write negatively about it.
- Please make the summary coherent and fluent in terms of sentence and information structure. Iterate over the written summary multiple times to improve it, and re-read the reviews whenever necessary.
- The summary should not look like a review, please write formally.
- Keep the length of the summary reasonably close to the average length of the reviews.
- Please try to write the summary using your own words instead of copying text directly from the reviews. Using the exact words from the reviews is allowed but do not copy more than 5 consecutive words from a review.

9.5 Human Evaluation Setup

To perform the human evaluation experiments described in Sec 5, we hired workers with 98% approval rate, 1000+ HITS, Location: USA, UK, Canada, and the maximum score on a qualification test that we had designed. The test was asking if the workers were native English speakers, and was verifying that they correctly understood the instructions of both the best-worst scaling and content support tasks.

9.6 Best-Worst Scaling Details

We performed human evaluation based on the Amazon and Yelp test sets using the AMT platform. We assigned workers to each tuple containing summaries from COPYCAT, our model, LEXRANK, and human annotators. Due to dataset size differences, we assigned 5 and 3 workers to each

1st	I bought this as a gift for my husband.
	I've been using Drakkar Noir Balm for over twenty years.
	I purchased these for my son as a kind of a joke.
2nd	This is the best product you can buy!
	You get what you pay for.
	Please do yourself a favor and avoid this product.
3rd	This is his every work day scent.
	It's very hard to buy the balm separately.
	It smells like Drakkar, but it is hard to find.
No Pronouns	Very nice, not too overpowering.
	This product has no smell what ever.
	Nice to use for hardwood floors.

Table 12: Examples of review sentences that contain only pronouns belonging to a specific class.

tuple in the Amazon and Yelp test sets, respectively. We presented the associated reviews in a random order and asked the workers to judge summaries using the Best-Worst scaling (BWS) (Louviere and Woodworth, 1991; Louviere et al., 2015) that is known to produce more reliable results than ranking scales (Kiritchenko and Moham-mad, 2016). The judgment criteria are presented below, where *non-redundancy* and *coherence* were taken from Dang (2005). *Fluency*: the summary sentences should be grammatically correct, easy to read and understand; *Coherence*: the summary should be well structured and well organized; *Non-redundancy*: there should be no unnecessary repetition in the summary; *Informativeness*: how much useful information about the product does the summary provide?; *Sentiment*: how well the sentiment of the summary agrees with the overall sentiment of the original reviews?

9.7 Points of View

Summaries differ from reviews in terms of the writing style. Specifically, reviews are predominantly

written informally, populated by pronouns such as *I* and *you*. In contrast, summaries are desirable to be written formally. In this work, we observed that a surprisingly simple way to achieve that is to condition the generator on the distribution over pronoun classes of the target review. We computed pronoun counts and produced the 4 class distributions: 1st, 2nd, 3rd person POV, and ‘other’ in case if no pronouns are present.

Consider the example sentences in Table 12. Here one can observe that the sentences of different pronoun classes differ in the style of writing and often the intention of the message: 1st POV sentences tend to provide clues about the personal experience of the user; 2nd POV sentences, on the other hand, commonly convey recommendations to a reader; 3rd POV and ‘other’ sentences often describe aspects and their associated opinions.

Gold	Bennett Medical has poor customer service. Phone calls can take a long time to get answered and leaving voice mails tend to be fruitless. The products are overpriced and take a long time to be refilled. Using this medical supply company can be a hassle.
Ours	This medical supply is the worst medical supply company in the valley. The customer service is horrible, the staff is rude, the wait times are long, and the service reps are not helpful at all. Do not recommend this company to anyone.
Copycat	If I could give them 0 stars I would. The customer service is terrible, the staff is extremely rude and helpful. If you're looking for a new provider, this is the place to be. I will never use them again.
MeanSum	Service is horrible, especially the manager. I have a lot of kids but not this place. Two months later I was able to go in and get a new one to go in the next day. They would tell me that they would have to do a lot of our water to be there to get a new one. But this is the first time I have dealt with him and we will never go back again. Thanks for your hard work, and I will never go anywhere else.
Lexrank	Bennett Medical for Cpap supplies are horrible. Never enough staff to answer phone, so you'll need to leave messages. DON'T use this medical supply. If I could give Bennett Medical zero stars I would! Will be moving to another medical supply as soon as I can.
Review 1	Bennett Medical for Cpap supplies are horrible. We have waited for three weeks to refill supplies and we are still waiting. This company does not have good customer service, you can only leave messages, and they never call back. If I could give Bennett Medical zero stars I would!
Review 2	Teachers Health Trust, please look into the practice of the billing and filling of durable services. The mask cushions go for 45 to 50 days because of the lack of communication. The people in charge of billing are very argumentative and lack customer service. I will drop them after annual, because of my insurance obligations.
Review 3	Fantastic service from Jocelyn at the front desk, we had a really hard time getting the right paperwork together from Drs but she stuck with us and helped us every step of the way, even calling to keep us updated and to update info we might have for her. Thanks Jocelyn.
Review 4	I hardly ever write reviews, but I'd like to spare someone else from what I experienced. So a warning to the wise... If you like rude incompetent employees, almost an hour long wait for just picking up a phone order, and basically being treated like a second class citizen then look no further than Bennett Medical.
Review 5	DON'T use this medical supply. Never enough staff to answer phone, so you'll need to leave messages. No return phone calls. I am unable to get my CPAP supplies every quarter without hours of calling / waiting / calling. Poor customer service. Will be moving to another medical supply as soon as I can.
Review 6	Terrible experience. They have ridiculous price also bad customer services. You can get nebulizer machine around \$50 at amazon, bennet medical charge you almost twice more expensive price. And breathing kit price was unbelievable too. Because of deduction, I had to pay them all out of my pocket whatever they charged. I don't recommend this medical company to anyone.
Review 7	Good luck getting a phone call back or someone to answer the phone without hanging up immediately. I have called over 20 times left 5 voicemails over the last 30 days, just to refill a mask perscription. This is an ongoing issue that is beyond frustrating. Not trying to destroy this businesses name just want the owners to implement some basic customer service skills.
Review 8	Always receive friendly customer service whenever we call or go into the location. My questions are always answered and I am very happy with the supplies we get from them. Great people providing a great service! Thank you for all you do!

Table 13: Example summaries produced by different systems on Yelp data.

Gold	It is very clean and nice inside. Everyone is so kind and friendly. They do an amazing job on both nails and pedis. They get it done with speed and precision with a price that is very much affordable. They have the best customer service.
Ours	This nail salon is very clean and the staff is very friendly. They have a wide variety of gel colors to choose from. The prices are very reasonable and they do a great job. The nail techs are very nice and do great work.
Copycat	This is the best nail salon I have ever been to. Everyone is very friendly and professional. My nails look great and I'm glad I did! I will definitely be coming back to this place.
MeanSum	The owner is so nice and accommodating. I went to get my nails done by a friend, and I was extremely happy with this salon. Everyone was very friendly and I was able to use them for nails. They did a great job on my nails and the best part about it was that it was a busy day but it was a treat! Highly recommend them.
Lexrank	I really enjoy coming here to get my nails done. B did an amazing job on my nails. Amazing service and nails. However B did an AMAZING job on my coffin chrome nails and Nancy was extremely helpful figuring out how I wanted my nails done too. Everyone is so friendly there too.
Review 1	Tim and Tami always always always have the best customer service and do the best nails. I will NEVER go anywhere else. Even after weeks my nails look and feel as good as they did when I first got them done! I'm so dedicated I recommend and bring in all my friends!
Review 2	Definitely my new nail salon! Everyone is so friendly and kind, I felt so welcomed! B did an amazing job on my nails. He made sure everything was perfect and happily changed something to make me happy. I would highly recommend this place to anyone who wants A + work at a totally affordable price. Love it!!!)
Review 3	Amazing service and nails. This is the second time I have been here, they did a perfect job again. They get it done fast yet with precision. Everyone is so friendly there too. Best nail salon I have ever been too. I'm glad I found it.
Review 4	I really enjoy coming here to get my nails done. They do a wonderful job on both pedis and nails. It is nice and clean inside. They are very friendly and welcoming. It is worth it to stop in and try it out.
Review 5	My first set of acrylics ever... I decided 27 years was a lot enough time to wait, and I'm SO happy with them. I'm not a huge nail person, and was glad to stumble upon this salon. My nail tech was quiet, clean, and very detail-oriented. Very pleased with my experience here and I recommend this place.
Review 6	I called to make an appointment for later today for 3 adults and 2 kids and the man who answered the phone said 'we only have 2 techs today' we can't do that. Poor customer service and I never even went in.
Review 7	Golden Nails has been my nail place for almost a year so it was surprising to see new management. However B did an AMAZING job on my coffin chrome nails and Nancy was extremely helpful figuring out how I wanted my nails done too. Definitely excited to keep coming back!
Review 8	Seriously the best service I have ever gotten at a Tempe nail salon!! I walked in and they helped me right away. Nancy helped me pick the perfect color and was very honest and up front about everything! I wanted something very natural and using the dip method, I love my nails!!

Table 14: Example summaries produced by different systems on Yelp data.

Gold	These are a very comfortable and cute sandal. This thong sandal goes with a variety of outfits and the cushy sole allows for all day comfort. However, they do run a little small, sizing up provides a better fit. Overall, a reasonably priced shoe that will last for years to come.
Ours	These sandals are very cute and comfortable. They fit true to size and are very comfortable to wear. They look great with a variety of outfits and can be dressed up or down depending on the occasion.
Copycat	I love these sandals.They are very comfortable and I can wear them all day without any discomfort. I wear them to work and they are comfortable to wear.
MeanSum	I love these shoes. They are so comfortable and I love the style. They are very comfortable and the perfect price! I would definitely recommend this product to anyone. They are comfortable and stylish.
Lexrank	I have been wearing White Mountain beaded sandals for a couple of years now and they are wonderful. I will never buy from white mountain again. I love White Mountain sandals. Lots of compliments every time I wear them.
Review 1	I get constant compliments on these sandals. I order them every summer in a variety of colors. I had heel spurs and back problems so the cushy softness of these is the only thing I can wear comfortably and the small wedge heel is perfect for my back.
Review 2	These thongs are fun, festive, flexible and surprisingly comfortable. I have very sensitive feet and I can wear these cuties all day. The arch support is great and there is a nice give in the sole. I love these so much I want to put a few pairs away in case they discontinue them. They go with everything.
Review 3	I have been wearing White Mountain beaded sandals for a couple of years now and they are wonderful. They are lightweight and cushion the feet when worn for long hours. They are also beautiful and usually hold up for two or more seasons.
Review 4	This was great price for this cute sandal. Unfortunately, the toe piece was very hard and they were a little narrow... unusual since I normally wear a B width. For the right person, they would probably be fine. They just didn't work for me.
Review 5	I love White Mountain sandals. this is my 2nd pair of these shoes. I wore out the last pair after 2yrs. They are very very blingy and I like that. Would I order another pair? you bet I would / will.
Review 6	Item was too small, purchased for a friend their size 9 is smaller than the size 8 in the store. Sent it to the wrong address, and I can not seem to find anyone that will tell me where my bill is. I will never buy from white mountain again.
Review 7	I lived in sandals that looked exactly like this but I thought they were by Bjorn. I couldn't find them anywhere, but found these (go figure). While they aren't quite as comfy as my other ones, I think with a bit of breaking in they'll be just fine. Lots of compliments every time I wear them.
Review 8	Not only are these SUPER comfortable (yes, even between your toes), they look great with just about anything I wear! I have been complimented on these daily!! I typically wear a 6 1 / 2, I ordered a 7 and they fit perfect!! I need more of these!! HIGHLY RECOMMENDED!!!

Table 15: Example summaries produced by different systems on Amazon data.

Gold	This is a perfect compact table that fits well in many places. The chairs are surprisingly very comfortable as well. It is cute and perfect for smaller living quarters and the best part is assembly is simple and straightforward.
Ours	This is a very nice table set for the price. It was easy to assemble and looks great in the kitchen. The only problem is that it is not sturdy enough to hold a lot of weight. It would be nice if it had a little more weight to it so that it would not tip over.
Copycat	This is a great table set for the price. It was easy to put together and looks great. The only thing is that the chairs are a little flimsy, but they are easy to assemble.
MeanSum	The table was very easy to assemble and was easy to assemble. The only thing I would say is that the box is very small and not very sturdy. The table is very sturdy. I would recommend it to anyone looking for a sturdy table and to put on the wall.
Lexrank	The table and chairs are very nice but not quite the color I expected (but I am getting used to it). The table and chairs are solid and sturdy! I received this table and chairs completely damaged. Table and chairs delivered by the carrier right on time and with no damage.
Review 1	It was easy to put together and looks great. However, when the item was shipped to me, one of the backs of the chairs was broken. I just fixed it myself with wood glue. Its not even visible now. The rest of it was in perfect condition.
Review 2	The table and chairs are very nice but not quite the color I expected (but I am getting used to it). Table and chairs delivered by the carrier right on time and with no damage. Very easy to assemble, but very difficult to get out of the box it was so well protected.
Review 3	This table was super easy to put together. The table and chairs are solid and sturdy! The seats are very comfortable. The table is the perfect size for our not so big kitchen. We are very pleased with this purchase.
Review 4	Moved to smaller living quarters and this just fits the bill. Color is perfect and it was easy to assemble. One fault to find is that the top scratches easily. It even came with a scratch. Other than that it is fine.
Review 5	I love my new dining room set. The set is very sturdy, the walnut finish is a nice color. This set is great for a small area, kitchen nook. Would not recommend for a large eating area. Table is small and so are the chairs. Yet strong enough to hold big boys and girls, thumps up, great price, packed well, arrived in a timely matter.
Review 6	It fits perfectly in the kitchen at the office. My staff assembled it without any delay. Everyone loves the dining set and they can't believe I ordered it on-line. I made the measurements and made sure of the dimensions of the room and the dining set and it's a perfect fit.
Review 7	I received this table and chairs completely damaged. The customer service experience with this company was terrible. In my opinion, this set is cheap and overpriced. It's not durable and not worth the money. Don't waste your time.
Review 8	The box looked like it had been opened, and then re-taped for resell. One of the chairs was broken, and the broken piece was nowhere close to the originating piece. Possibly other pieces damaged too, though didn't bother looking, instead just re-taped it back up to be sent back. I hope they don't just resell it to someone else.

Table 16: Example summaries produced by different systems on Amazon data.