Towards an Automated Writing Assistant for Online Reviews

Parisa Kaghazgaran
Texas A&M University
College Station, TX, USA
kaghazgaran@tamu.edu

James Caverlee
Texas A&M University
College Station, TX, USA
caverlee@tamu.edu

Workshop proceedings *Automation Experience across Domains*
In conjunction with CHI'20, April 26th, 2020, Honolulu, HI, USA
Website: http://everyday-automation.tech-experience.at

Abstract
Online reviews play a critical role in persuading or dissuading users when making purchase decisions. And yet very few users take the time to write helpful reviews. Encouragingly, recent advances in AI offer good potential to produce review-like natural language content. However, The main challenge is a lack of large, high-quality labeled data at both the aspect and sentiment level for training the automated review generators. Hence, we study the feasibility of a writing assistant framework in order to help users post online reviews and introduce a data-driven approach to label data required for training. We evaluate the effectiveness of our approach by launching a user study on how end-users perceive the quality of the labels and automated generated reviews.

Author Keywords
Intelligibility; review generation; review writing assistant; user-based evaluation

Introduction
There is a growing attention in creating new methods to help users share their opinions on review platforms. For example, Airbnb site require hosts and guests to write mutual reviews [3]. However, such a requirement may be an impediment to customer engagement in other platforms that feature products like movies and books. In another

promising direction, new tools based on natural language generation have shown good success in some domains. For example, carefully configured templates can transform well-structured data into legible text, especially for domains with consistent format and structure like weather forecast reports [1], Olympics stories [6], and corporate earnings reports [5]. However, such methods face challenges for online reviews that typically cover a broad range of categories (e.g., apps, products, restaurants) with multiple aspects within each category (e.g., food, service, staff and so on in restaurant domain) and diverse opinions that do not fit single template.

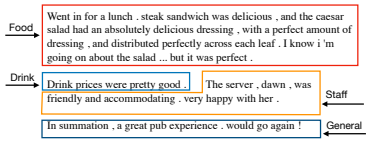

Figure 1: Example of Review Segmentation. Users express their opinions on multiple aspects of the item in a single review.

ent advances in AI offer good potential to produce ew-like natural language content [7]. Learning a deep ral review generator requires large amounts of labeled a. While there are many existing collections of online ews, very few have labels at the granularity of *aspects* price, food quality, or decor) and with *sentiment* associd with these aspects. Furthermore, it is unclear if incorating such aspect and sentiment into a review generator would result in meaningful reviews. To overcome these challenges, this paper introduces a data-driven approach to expand a small set of labeled reviews in order to obtain a large amount of labeled reviews demanded by neural networks and study the feasibility of a writing assistant framework in order to help users post online reviews. We propose automation at both review labeling and review generation levels and evaluate the quality of the labels and generated reviews through a user-based study.

Seed data [4]. It is used to label the unlabeled reviews containing 1,472 pairs of reviews and their labels.

Yelp Dataset. It is used to train the review generator containing about 4M reviews on 60k restaurants.

Automatic Review Labeling

Since users express opinion on different aspects of the target in a single review, a review cannot be labeled in whole to state a specific aspect (see Figure 1). We first split a review into its topically coherent segments and propose to label the resulting segments.

Review Segmentation.

The review segmentation algorithm traverses through each review sentence by sentence in order to cluster coherent sentences into one segment. At its core, review segmentation is based on a sliding window technique with a window size of two. Each sentence is compared with the right most sentence in the previous segment and if their distance is less than a specific threshold τ, then the sentence is added to the segment, otherwise it forms a new segment. This process continues until the end of the review in linear time. We adopt the Word Mover's Distance (WMD) [2] to measure the similarity between sequential sentences. Rather than relying on keyword matching, it attempts to find an optimal transformation from one sentence to another sentence in the *word embedding space*. Briefly, word embedding technique map each word into a numeric vector such that co-occurred words are close in the new vector space.

Label Assignment.

The label assignment algorithm is based on a small seed set. The main intuition is to find semantically similar seeds to the unlabeled segments and use their labels to identify the aspect of the segments. For this purpose, we compare each sample in the seed set against the unlabeled segments using the WMD distance function. The unlabelled segment receives the label of closest seed sample.

Table 1 demonstrates examples of segments obtained from the segmentation algorithm and their labels assigned by the label assignment algorithm across different aspects and

Table 1: Example of segments obtained from the proposed segmentation algorithm and their corresponding labels obtained from the label assignment algorithm across different aspects and sentiments. (+) and (-) indicate positive and negative sentiments, respectively

Aspect-specific Review Segments	Label
steak sandwich was delicious and the caesar salad had an absolutely delicious dressing with a perfect amount of dressing and distributed perfectly across each leaf . i know i m going on about the salad .	**Food (+)**
today was my second visit to the place after having a good first experience but i am so disappointed with the quality of the food that i can say it has been my worst experience of food in months the sun dried tomatoes very absolutely stale to an extent that they tasted bitter the pizza base was so thick that it was uncooked and soggy the four cheese blend tasted completely different than the last time and so did the pesto sauce . no consistency with food quality .	**Food (-)**
the ambiance is nice too . it s a bit dark but they have this nice light display above on the ceiling made with mason jars . there is a comfy seating area in the bar area that s nice too .	**Ambience (+)**
however the one thing that surprised me was how dirty the restroom was in this restaurant . the floor was really dirty and toilet papers were unwell kept . the restaurant could at least have someone maintained the restroom in good shape and clean because this will reflect on how one maintains the cleanliness of the place .	**Ambience (-)**
the price is very reasonable for a family of four with plenty of leftovers to take home .	**Price (+)**
my wife i had a groupon for this place and for the price it was very poor value quality .	**Price (-)**
i had a nice glass of california cabernet . the wine list while not expansive was good . the bartender i had seemed to have a nice knowledge of what was going on with the wine that encompassed it .	**Drink (+)**
i ordered a glass of Merlot that was delivered to me in a dirty glass . the waitress was very polite and went to get me a new glass of wine but i was still unimpressed at that point .	**Drink (-)**
highly recommend for lunch . even during lunch rush it was not super packed . this would be a good place for a lunch meeting .	**General (+)**
i am not sure why anyone would like this place . the only thing it has going is location and that is simply not enough not for me .	**General (-)**

sentiments.

Evaluation.
We evaluate if automated labels are comparable with manual labeling. We set up a crowd-based user study to verify if the labels are assigned truthfully according to human readers. We post 100 surveys on Amazon Mechanical Turk (AMT) each including a guideline and a set of reviews for which we seek a label from Turkers. The guideline has two major points: (i) it shows a sample of reviews along with their labels from the seed set to provide a context on how reviews and labels are paired with each other, (ii) it asks

Label	Accuracy(%)
Food (+)	94.33
Food (-)	85.66
General (+)	87.66
General (-)	81.00
Ambience (+)	81.33
Ambience (-)	77.66
Price (+)	86.00
Price (-)	68.00
Drink (+)	82.23
Drink (-)	57.66

Table 2: Majority of the automated labels are recognized as accurate by human evaluators (> 80% acc).

Label	Accuracy(%)
Food (+)	93.07
Food (-)	97.69
General (+)	86.15
General (-)	85.38
Ambience (+)	97.69
Ambience (-)	90.76
Price (+)	90.00
Price (-)	83.07
Drink (+)	90.76
Drink (-)	33.84

Table 3: Majority of the generated reviews are perceived as reliable by human evaluators (90% acc).

Turkers to label the reviews through a series of multi-choice questions. We design 100 surveys each with 10 reviews to cover all the labels. Each unique survey is assigned to three workers, i.e., 3 HITs (Human Intelligence Task) per task, giving us a total of 300 surveys and 3,000 questions.

To ensure the quality of responses, we insert a trivial question into each survey, which asks the Turker to check if a mathematical equation is False or True. It helps to manage the risk of blindly answered surveys. Furthermore, we only accept surveys from Turkers with approval rating of at least 95% and those who dwell on the survey for at least 7 minutes. We also restrict our tasks to workers located in the United States to guarantee English literacy.

Table 2 demonstrates the performance of automatic labeling against human judgment across various labels. The majority of the labels are found accurate by human evaluators with at least 80% accuracy. However, the accuracy for labels Price/negative and Drink/negative is relatively low and we can relate this to the fact that these labels do not have a significant representation in the seed set.

Automatic Review Generation

Here, we introduce a generative model based on language neural networks to automate user review generation. The main intuition is to propose a sequence to sequence model where the first sequence encodes the labels of the segments obtained from previous step into a context vector. This context vector along with review words are the input to the second sequence in order to train the review generator. Formally, Given the aspect A as input attribute, we aim to generate a review sequence $R = (w_1, .., w_{|R|})$ constrained by the input to maximize the conditional probability $p(R|A)$:

$$p(R|A) = \prod_{t=1}^{|R|} p(w_t|w_{<t}, A) \quad (1)$$

where $w_{<t}$ refers to the tokens seen until time step t.

Evaluation.
Similar to label assessment, we launch a crowd-based user study by posting surveys on AMT. We follow similar guidelines to ensure the quality of the answers. We design 100 surveys each with 10 generated reviews at various aspects. We assign 3 HITs per task, giving us a total of 300 surveys and 3000 questions. We ask Turkers to label the model-generated reviews through a multi-choice questions based on the aspect.

From Table 3, we observe that generated reviews stay with the desired aspect with higher than 90% accuracy for a majority of the labels. For example, 93% and 97% of reviews on Food (positive and negative) are perceived equally by the model and the human evaluators while this number is 34% for drink/negative. We can relate this to the fact that the label *Food* has a better representation in both seed set and our expanded dataset as it is the main topic of discussion when writing a review for a restaurant.

Conclusion and Discussion

We have explored how to make use of automation to help users writing online reviews in particular when sufficient data required by neural networks are not available. We automate the review writing process at two steps: (i) build a ground truth of reviews at aspect and sentiment level and evaluate the effectiveness of the labels by human readers, (ii) propose a generative model that produces reviews conditioned on input aspects and evaluate the quality of the generated reviews through user study. In the next step, we

aim to study how users are willing to use the proposed system and how we can incorporate their intention at design level.

REFERENCES

[1] Ibrahim Adeyanju. 2015. Generating weather forecast texts with case based reasoning. *arXiv* (2015).

[2] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*.

[3] Dean D Lehr. 2015. An analysis of the changing competitive landscape in the hotel industry regarding Airbnb. (2015).

[4] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, and others. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 19–30.

[5] Paul Roetzer. 2016. How the Associate Press and the Orlando Magic Write Thousands of Content Pieces in Seconds, https://bit.ly/2HCyAiS, Last Access: 08/14/2019. (2016).

[6] WashPostPR. 2016. The Washington Post experiments with automated storytelling to help power 2016 Rio Olympics coverage, https://wapo.st/2G67Sg6, Last Access: 08/14/2019. (2016).

[7] Yuanshun Yao and et al. 2017. Automated crowdturfing attacks and defenses in online review systems. In *CCS*.