Mining and Organizing User-Generated Content to Identify Attributes and Attribute Levels

Artem Timoshenko

MIT Sloan School of Management
Massachusetts Institute of Technology

John R. Hauser

MIT Sloan School of Management
Massachusetts Institute of Technology

Abstract

We investigate User-Generated Content (UGC) as a source of customer needs from which to identify attributes and attribute levels for a high-craft conjoint-analysis study. Non-informative and repetitive content crowd out information about customer needs in a large corpus of UGC. We design a machine-learning hybrid approach to enhance customer-need extraction making it more effective and efficient. We use a convolutional neural network (CNN) to identify informative content. Using pre-trained word embeddings, we create numerical sentence representations to capture the semantic meaning of UGC sentences. We cluster sentence representations and sample sentences from different clusters to enhance the diversity of the content selected for manual review. The final extraction of customer needs from informative diverse sentences relies on human effort. In a proof-of-concept application to oral care, we compare customer needs identified from UGC to customer needs identified from experiential interviews. First, our analyses suggest that, for comparable human effort, UGC allows identifying a comparable set of customer needs. Second, machine learning enables analysts to identify the same number of customer needs with less effort.

Motivation

A conjoint-analysis study is only as good as the attributes upon which the study is based. Missing important attributes lowers the quality of the study and leads to inefficient product development. Identifying new highly-valued attributes and attribute levels leads to major breakthroughs in product strategy. Consider "Attack," a laundry detergent introduced by the Kao Group in the 1980s. At the time, the customer needs for laundry detergents were well established: cleaning, safe and gentle, good for the environment, ready to wear after drying, easy to use, smell fresh and clean, and value. To design new detergents, most manufacturers focused on combining attributes to address these customer needs. Perceived "value" played a major role in the market for detergents. For example, detergents were sold in large "high-value" boxes to enhance perceived value. Figure 1 compares a vintage Tide box with Attack's packaging at its launch.

Kao did not limit itself to established attributes and attribute levels. Japanese consumers did not have the space to store laundry detergent in their apartments and, as a result, they went to the store often. Consumers commonly went by bicycle or by foot. Kao recognized an unmet customer need and the corresponding attribute level (the need for small package for the same cleaning power). Kao launched Attack, a highly-concentrated detergent in an easy-to-store and easy-to-carry package. Laundry customers were willing to pay a substantial price premium for this product and, within a year, despite the higher price, Attack commanded almost 50% of the Japanese laundry market (Kao Group 2016). Other firms, including US-based firms, were slow to identify this customer need and did not immediately include the 'low-package-size' attribute level in their marketing studies, which gave Kao a competitive advantage.

Figure 1. Vintage Tide Detergent Box and Attack's Package at Launch[1]

Examples of successful major innovations based on newly identified attributes and underlying customer needs include the touchscreen features in the smartphone category and Procter & Gamble's Swiffer mop (Continuum 2016). Even point-of-care blood-gas testing in intensive care units of hospitals was revolutionized when the need for new attributes for these important medical instruments was recognized, analyzed, and satisfied.

These examples come from product development, but conjoint analysis is also used widely to value patents and copyrights (Cameron, Cragg, and McFadden 2013). Accepted litigation practice pairs a marketing expert, who provides estimates of willingness to pay, with a "damages" expert, who handles the implications of WTP. The damages expert testifies about the value of the

[1] Tide image from https://www.pinterest.com/blacklab3/vintage-soap/. Attack image from http://www.kao.com/group/en/group/history_01.html.

patent or copyright. Recently, Allenby et al. (2014) proposed that the marketing expert play both roles. Instead of computing WTP, the authors propose that conjoint analysis be used directly to estimate the change in market price that is due to the patent. They propose that a conjoint-analysis simulator be used to determine the (Nash) market equilibrium prices at which all firms in the market simultaneously select maximum-profit prices, each assuming the other firms do not change their prices. Their proposed method requires a reasonably complete set of attributes, because equilibrium prices depend upon the error term in conjoint analysis which, in turn, depends on unmodeled attributes. See Eggers, Hauser, and Selove (2016) in this volume. The courts have intuited this dependence. When conjoint analysis is used for more than WTP (or willingness to buy, WTB), some courts have disallowed testimony from conjoint-analysis experts because the courts perceive that the attribute description is inadequate (e.g., Alsup 2012).

Whether a conjoint analysis is used to price a product, identify new product opportunities, estimate the impact of a change in attributes, or value copyrights and patents, it is important that the conjoint-analysis study is based on a rich set of attributes for the product category. The accuracy and relevance of the conjoint-analysis study depends on the quality and completeness of the attribute-based description.

Typical Approaches to Identify Attributes

Direct Approaches. Often, the client provides a list of attributes and attribute levels and asks the analyst to design and execute the conjoint-analysis experiment. This is a perfectly fine approach, but pushes the responsibility back to the client to specify an appropriate list of attributes. Alternatively, an analyst might search competitive websites, search websites that compare and contrast products, and search websites that make recommendations. Advertising claims complement these Internet searches. If the market is relatively stable, or if the conjoint analysis is used for WTP or WTB, then well-conducted Internet searches are an efficient means to identify the attributes for the conjoint-analysis study. Internet searches are less useful if the market is in flux, or if the goal is to identify new innovations. "Unarticulated" needs and attributes might not be found in these Internet searches because no existing product has the attributes. New opportunities could be missed. Analysts must also be careful because comparison websites focus on points of difference among products. They might miss basic "must have" attributes.

Indirect Customer-Based Approaches. Indirect customer-based approaches begin directly with the customer. Focus groups and experiential interviews enable customers to articulate their needs and desires for the product category. The analyst experiences the experiences of the customers. Rather than asking directly about attributes, the analyst seeks first to understand the customers' needs and then translates those customer needs into attributes (solutions) that address the customer's expressed needs. Fortunately, there are a variety of proven methods to translate customer needs into attributes, including hedonic regression, Quality Function Deployment, and the Brunswik "lens" model (Brunswik 1952; Hauser and Clausing 1988; Sullivan 1986).

The direct approaches are easier to implement and less expensive, but the indirect customer-need-based approaches provide certain advantages. Indirect approaches identify a broad set of attributes with less functional overlap. This is particularly valuable because survey formats and respondent attention often limit the number of attributes and attribute levels. Furthermore, indirect approaches often identify unmet customer needs that lead to successful innovations.

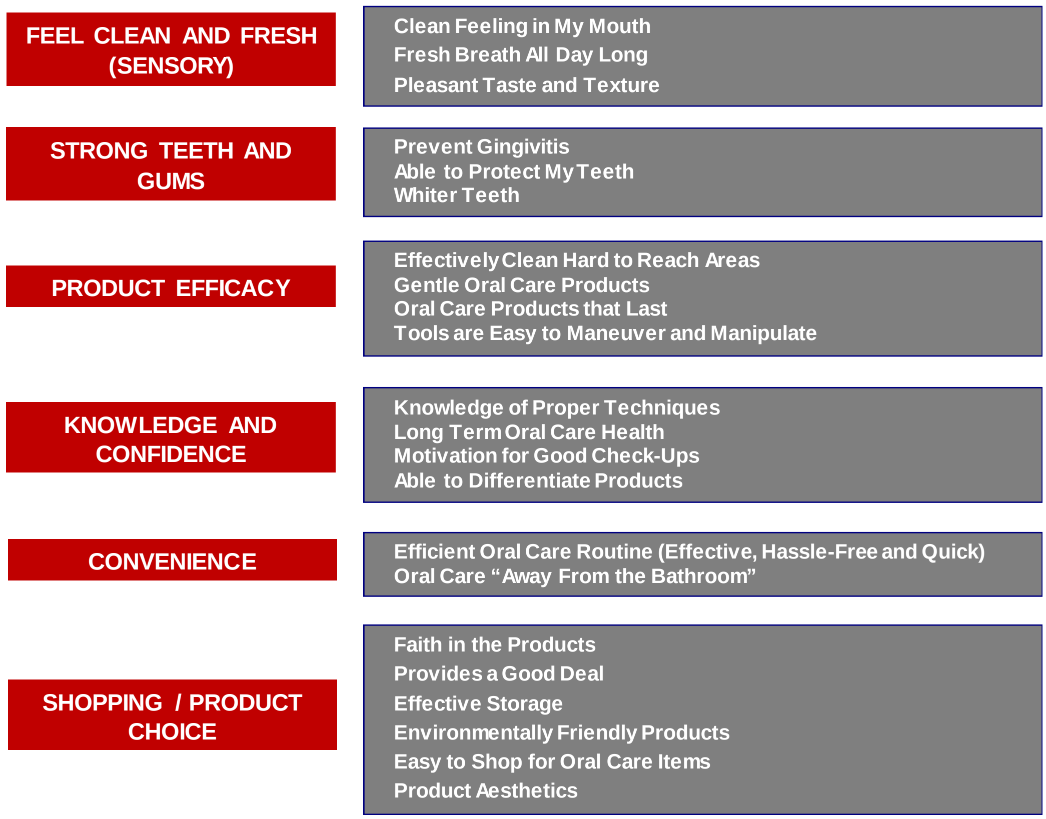
Our study focuses on identifying customer needs for an indirect approach. We rely on established methods to translate customer needs into attributes and attribute levels.

CUSTOMER NEEDS VERSUS CUSTOMER SOLUTIONS. Customer needs, as used in this paper, are abstract statements that describe what a customer seeks to obtain from a product in the category. For example, in oral care, a customer need might be: "Easy to know the correct amount of mouthwash to use." Customer needs are purposefully abstract so that they provide sufficient flexibility for the firm to design attributes that fulfill customer needs. With these definitions, attributes in conjoint analysis are solutions to customer needs. For example, a solution to the customer need might be to put "ticks on a cap that is used for dosage" or "pictures and numbers on the bottle to indicate dosage." See Figure 2.

Figure 2. Attribute-based Solution to Customer Need to Know Easily the Correct Amount of Mouthwash to Use

VOICE OF THE CUSTOMER. A structured set of customer needs is often called the "voice of the customer (VOC)." The most common VOC method consists of four steps: (1) experiential interviews with customers, (2) sentences highlighted by multiple human judges, (3) "winnowing" to obtain a reduced, non-redundant set of customer needs, and (4) methods to organize the customer needs into an hierarchy of "primary," "secondary," and "tertiary" customer needs (Ulrich and Eppinger 2016; Griffin and Hauser 1993; Herrmann, Huber, and Braunstein 2000). There are at least two common procedures to organize customer needs into an hierarchy: (1) affinity groups where customers, themselves, sort the needs, and (2) card-sort methods where customers sort together customer needs that are similar and analysts cluster customer-need co-occurrence matrices. Figure 3 provides an example of the first two levels of a customer-need hierarchy that was delivered to an oral-care client. This VOC was produced by a marketing consulting firm with almost thirty years of experience in the voice of the customer.

Figure 3. Voice of the Customer for Oral Care Products



User Generated Content (UGC)

User-generated content (UGC) is text (and pictorial) content about products that customers themselves generate. For example, Twitter posts, customer blogs, and customer reviews are all UGC. UGC might also include customer-complaint data or data collected from customer-help records. UGC is an exciting new source of information from which customer needs (and conjoint-analysis attributes) can be extracted. UGC is often available quickly and at low incremental cost to the firm. UGC is updated automatically and never gets stale.

However, UGC presents its own challenges. First, there are often too much data for human readers to process. For example, there are over 115,000 oral-care reviews on Amazon consisting of over 400,000 sentences. Human readers just cannot process that entire corpus. Second, much of the data in UGC are repetitive and not relevant. Sentences such as "I recommend Crest for oral care." do not express any customer need. We expect, and our analysis confirms, that most of the UGC on oral care concentrates on a relatively few needs. Third, UGC data are unstructured and mostly-text based. Identifying customer needs requires a thorough understanding of the content, and the unstructured nature of UGC complicates automatic analysis.

Our Goals

UGC versus Experiential Interviews. Our first goal is to compare in completeness and quality a set of customer needs, identified from UGC, to customer needs identified by standard methods as practiced by experienced analysts working from high-quality experiential interviews. Ideally, UGC-based customer needs should (1) have a substantial overlap with interview-based customer needs, (2) miss relatively few interview-based customer needs when limited to comparable analyst effort, and (3) include customer needs that were not identified from an exhaustive search of experiential-interview transcripts. We feel that if we confirm that customer needs from UGC satisfy these criteria then we validate UGC as a viable replacement for costly experiential interviews.

Machine-human Hybrid versus Human-only Processing. Our second goal is to use machine learning (deep learning) to streamline the identification of customer needs from UGC. In particular, we seek to use machine learning to eliminate non-relevant content and organize the remaining content to minimize redundancy.

For example, suppose that analysts, who are experienced in the use of VOC methods, have the capability of reading N sentences from UGC to identify customer needs. (Their capability might be limited by time, monetary budgets, or simply attention.) Not all N sentences will be relevant and many sentences will describe redundant customer needs. Let's suppose that the analysts can identify K_o unique customer needs. A machine-human approach is more efficient if it can identify at least K_o customer needs with human effort that is less than or equal to that which would have been required for VOC experts to review N random sentences and identify K_o customer needs. (Computational costs are trivial compared to human effort.)

If we demonstrate that the machine-human hybrid is more efficient, then, with continuous improvement through application, the evolved method might be able to optimize the machine-human hybrid and achieve the best human-effort cost per identified customer need. (We assume that the machine-learning method is fully programmed. The computation cost is a very small fraction of human effort.)

Optimization of Human Effort. In our scheme, there are multiple types of human effort that enter any analysis. In standard VOC methods, experiential interviews are extremely costly. UGC eliminates recruiting, interviewing, and transcription costs. In the machine-human hybrid method, there are two types of human effort required. Human analysts review sentences to determine whether the sentences are "informative" or "non-informative." Then, human analysts review informative sentences to extract customer needs. The former is less onerous and time-consuming than the latter. For the purposes of this paper, we leave the optimization of human effort to future research. Such optimization requires that we quantify the value of additional customer needs and quantify the effort costs of interviewing, customer-need extraction, and informative-vs.-non-informative classification.

A PROPOSED MACHINE-HUMAN HYBRID FOR ATTRIBUTE IDENTIFICATION

WHY A MACHINE-HUMAN HYBRID. When machine-learning methods improve, we might be able to automate all stages in the identification of customer needs from UGC. To date, the final stage has defied automation. Formulations of customer needs must be precise for subsequent analyses. Moreover, the machine-learning methods are not sufficiently sensitive to semantic context to extract abstract customer needs from informative content. UGC is unstructured and not necessarily generated to articulate customer needs. Context matters and customer needs appear to be more than "buckets of words." For example, bucket-of-word methods, such as Latent Dirichlet Allocation (LDA, Blei, Ng, and Jordan 2003) and LDA with hidden Markov models (LDA-HMM, Griffiths, et al. 2004) do not seem to capture the semantic context necessary for identifying customer needs. But stay tuned.

We have successfully automated two critical tasks in the analysis of UGC: identifying informative content and sampling representative a diverse set of content for review. The resulting machine-human hybrid is more efficient, and equally as effective, as a pure human-effort-based method. We feel this is substantial progress in a relatively short time. Analysts have had almost thirty years of continuous improvement to optimize human-effort-based VOC methods.[2] VOC identification by experienced analysts is a challenging benchmark.

OVERVIEW OF THE MACHINE-HUMAN HYBRID. Table 1 provides an overview of the four stages in our proposed method. The stages are:

1. ***UGC***. Rather than relying on expensive experiential interviews, we harvest readily available UGC from either public sources or propriety company databases.

2. ***IDENTIFY INFORMATIVE CONTENT***. We use a machine-learning classifier called a convolutional neural network (CNN) to filter out non-informative sentences so that the remaining corpus is rich in informative content. Because a CNN is a supervised learning method, it must be "trained." Training requires human effort to classify a subset of sentences as informative vs. non-informative. In practice, the number of training sentences should be a small fraction of the corpus.

3. ***SAMPLE DIVERSE CONTENT***. We cluster "sentence representations" to select a set of sentences likely to represent diverse customer needs. Sentence representations are, in turn, based on

[2] Consulting firms, with experience in VOC methods, make human effort more efficient with software that makes it easy to highlight phrases in interview transcripts. Additional "bookkeeping-like" software makes it easy to keep track of redundant phrases during the winnowing process. Such proprietary software does not have the capabilities to be called machine-learning. These firms have also optimized human effort through training and experience.

dense numerical representations of words that capture semantic meanings.

4. *FINAL EXTRACTION OF REPRESENTATIVE CUSTOMER NEEDS.* Analysts review the winnowed, informative sentences to identify customer needs. In the machine-human hybrid approach, this final stage is based on human effort and is the same task as that used in existing human-effort-based methods.

We now describe the two machine-learning methods that we customized to the identification of customer needs. We then describe a proof-of-concept application to oral care.

Table 1. Automating Attribute Identification—Machine-Human Hybrid

Traditional	Machine-learning Hybrid
Experiential interviews	User Generated Content
Highlight informative sentences manually	Machine learning (convolutional neural network, CNN) identifies informative sentences
Reduce customer-need redundancy manually (winnowing)	Cluster numerical "sentence representations" to remove sentence redundancy and thus identify diverse customer needs
Extract customer needs manually from interview-based sentences	Extract customer needs manually from informative diverse UGC sentences

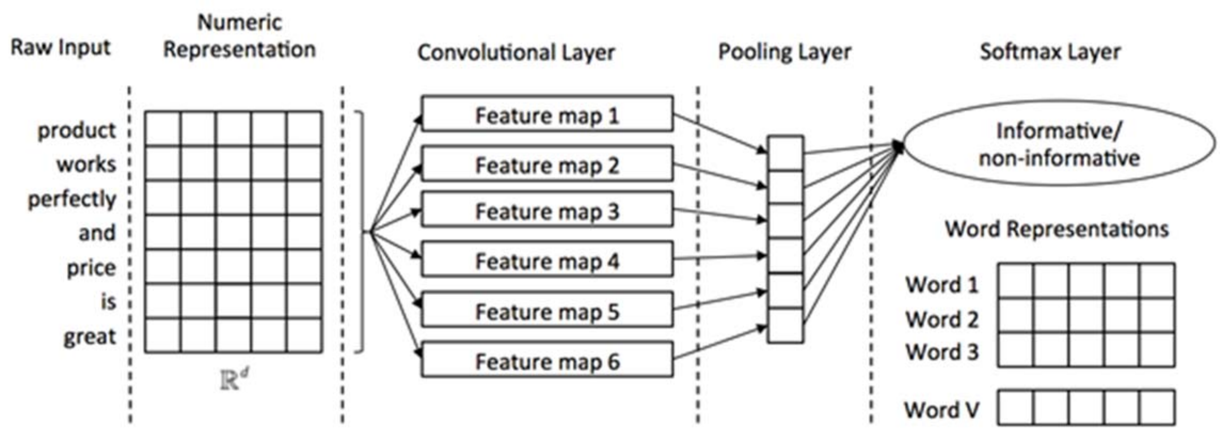
PREPROCESSING UGC TO IDENTIFY SENTENCES WITHIN THE UGC

Sentences are most likely to contain customer needs and are a natural unit by which human analysts process either experiential interviews or UGC. But in UGC, customers do not always use a sentence structure. We preprocess raw UGC to transform the UGC corpus into a set of sentences. We use an unsupervised sentence tokenizer from the natural language toolkit (Kiss and Strunk 2006). We automatically eliminate stop-words (e.g., 'the' and 'and') and non-alphanumeric symbols (e.g., question marks and apostrophes). We transform numbers into number signs and letters to lower case. We further screen sentences to account for the artifacts of grammatical or punctuation errors in UGC. In particular, we drop sentences that are too short (less than three words) or too long (more than ten words). UGC tends to have many fewer compound sentences than experiential-interview transcripts.

CONVOLUTIONAL NEURAL NETWORK (CNN)

We use a convolutional neural network (CNN) on the corpus of sentences after preprocessing to classify sentences as either informative or non-informative. A CNN is a supervised classification model (e.g., Kim 2014). We use a CNN to transform numerical representations of sentences into a prediction of whether or not the sentence is informative. A CNN has multiple types of layers and can have multiple layers of each type. Figure 4 illustrates the types of layers that are contained in our CNN. (Our CNN is not an off-the-shelf CNN, but rather customized for our application.) The two key properties of the CNN are that (1) the CNN learns how to quantify and classify sentences simultaneously in the model, and (2) the model is able to process input (sentences) of different length.

Figure 4. Examples of the Types of Layers in our Convolutional Neural Network



NUMERIC REPRESENTATIONS OF WORDS. For every word in the English-language dictionary, the CNN represents the word by a numerical vector. We use pre-trained 300-dimensional word embeddings as described in the next section. If the word embeddings were unavailable, and with sufficient training data, a CNN could be used to learn word representations simultaneously with other parameters. The CNN quantifies the sentence by concatenating the representations of the words.

CONVOLUTIONAL LAYERS. A convolutional layer begins by applying filters to the sentence representation. A filter selects varying contiguous subsets of the sentence representations and weights the elements of the subset. The CNN then applies non-linear transformations, such as a logistic function, to the weighted subsets. The result of the application of this transformation to various parts of the sentence representation is called a "feature map."

We calibrate the weights used in the filters by training the CNN on the sentences that have been coded by human effort. The number of filters and their sizes are hyperparameters of the CNN. We select these hyperparameters before the CNN is trained. We tune the hyperparameters with cross-validation.

Pooling layers. Convolutional layers often require many parameters and can become too complex to calibrate. If multiple convolutional layers are stacked without any dimensional reduction, then the number of parameters explodes. (The number of parameters also explodes if there are too many feature maps.) To maintain a feasible number of parameters, CNNs use pooling layers, which transform feature maps into shorter vectors. We use a max-pooling-over-time layer in which we retain the largest feature from each feature map produced by a convolutional layer (Collobert, et al. 2011).

Softmax layer. The final layer, called a softmax layer, in the CNN transforms the output of the final pooling layer into a prediction of whether the sentence is informative ($y = 1$) or not informative ($y = 0$). The softmax layer is a binary logit model applied to the output of the last pooling layer. The parameters of the logit model are calibrated with the training data. In our application, we assign a sentence as informative if the estimated probability is greater than 50%. Future applications might assign sentences to categories for further review based on other criteria.

Number of each type of layer. In our study, we stacked 3 convolutional layers and 1 pooling layer to generate input for the softmax layer. Each convolutional layer generates 40 feature maps. Performance of the trained CNN depends on a particular combination of layers and on the number of feature maps in convolutional layers. We used cross-validation to select these characteristics of the model.

CNNs vs. SVMs. Readers may be familiar with the use of support-vector machines (SVMs) for classification. CNNs have an advantage relative to SVMs because CNNs automatically and endogenously identify feature maps. In contrast, an SVM depends critically of the quality of the features used in the SVM. SVM features are often handcrafted, specific to application, dependent on context, and require substantial human effort. CNNs provide comparable performance to handcrafted SVMs without this substantial application-specific human effort (Kim 2014).

Clustering Sentence Representations

Armed with a corpus of informative sentences, we use machine learning to reduce redundancy. We cluster sentences that have similar semantic meaning and then sample from each cluster in proportion to the size of the cluster. For a given number of sentences, redundancy-reduced sentences are more likely to contain diverse needs than a random sample of informative sentences. Because the clustered corpus is designed for maximum diversity, it is more likely (for a given N) to yield a complete set of customer needs.

In order to cluster sentences, we create numerical representations of the sentences that capture semantic meaning. The transformation for clustering is different than the concatenation for

CNN classification, but both transformations are based on machine-language constructs known as "word embeddings." We first describe word embeddings and then describe how we aggregate word embeddings to sentence representations.

Word Embeddings. Word embeddings are the numeric vectors that capture the semantic meaning of words. The basic concept is that semantically similar words appear in similar contexts. Information about the contexts is then used to represent words in the numerical space. We rely on a high-quality pre-trained set of word embeddings that have remarkable properties. For example, if a word embedding, $v(w_i)$, is a vector representation of word w_i, then the $v(w_i)$ have the following properties (Mikolov et al. 2013a):

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen})$$

$$v(\text{walking}) - v(\text{swimming}) + v(\text{swam}) \approx v(\text{walked})$$

$$v(\text{Paris}) - v(\text{France}) + v(\text{Italy}) \approx v(\text{Rome})$$

We use 300-dimensional word embeddings that were pre-trained on the Google News Corpus using the "Skip-gram" model (Mikolov et al. 2013b). The Skip-gram model trains word embeddings by maximizing the average log-likelihood of words appearing within c words of one another in a sequence. For our purposes we simply adopt the word embeddings without further transformation.

Sentence representations. In the CNN we concatenated word embeddings. This operation matches the use of filters in the feature maps. To create sentence representations for clustering we use an operation that retains the centrality of the semantic meaning. For our proof-of-concept application in oral care, we adopt the averaging method advocated by Iyyer et al. (2015). This operation is based on machine-learning experience. For example, Iyyer et al. (2015) demonstrate that the average of word embeddings is as effective as explicitly modeling semantic and syntactic structure with neural networks or training sentence representations simultaneously with word embeddings (Le and Mikolov 2014; Tai, Socher, and Manning 2015).

Clustering sentence representations. Because sentence representations have the property that similar vectors represent sentences with similar semantic meanings, we cluster the sentence representations based on the Euclidean-distance norm. To be consistent with the hierarchical structures used in established VOC methods, we use an hierarchical clustering algorithm. Griffin and Hauser (1993) suggest Ward's method, which we adopt. Not only has Ward's method become standard practice for analyzing co-occurrence data, but, by using Ward's method, we maintain comparability with the human-effort-based benchmarks that we compare to the machine-human hybrid approach.

Final Extraction of Customer Needs. The clustered sentence representations, sampled proportional to size, provide a set of informative sentences that are designed to be rich in diverse customer needs. The final stage relies on trained analysts to read each sentence and extract the customer needs. We expect that human-effort extraction is more efficient with informative, diverse sentences than with sentences sampled randomly from the UGC corpus.

Oral Care Proof-of-Concept, Evaluation, and Comparison to Established Methods

We have three goals.

- Demonstrate that the machine-learning hybrid is feasible and that it can generate a set of customer needs from which attributes can be identified.
- Compare the relative customer-need content of UGC and experiential interviews.
- Evaluate the efficiency of the machine-human hybrid vs. a human-effort-based approach.

We select the oral care category because oral care is best described by a relatively broad and challenging set of customer needs, but the set of tertiary customer needs in oral care is not too large to make the analysis unwieldy.

"Gold standard" Human-Based Approach. A professional marketing consulting firm shared with us a VOC that they had delivered successfully to a client. Review Figure 3. The VOC was based on experiential interviews, with sentences highlighted by human analysts aided by the firm's proprietary software. After winnowing, customer needs were clustered by an affinity group. The output was six primary customer needs and 22 secondary customer needs (Figure 3), as well as further elaboration into 86 tertiary customer needs.

UGC Data. We consider 115,099 oral-care reviews from Amazon.com spanning the period from 1996 to 2014. Preprocessing with the sentence tokenizer produced 408,375 sentences.

Unique Dataset. To compare the customer-need information in UGC to the customer-need information in experiential interviews, we randomly selected 8,000 sentences from the UGC corpus. The sentence structure of UGC differs from that in experiential interviews. UGC sentences tend to be shorter and less compound. In experiential interviews, sentences tend to ramble as they do in normal conversation. They are not always complete, but make sense in context. Also, the questions asked by interviewers are part of the give-and-take and cannot be ignored. To affect a valid comparison, we asked analysts, with experience extracting needs from interview transcripts, to estimate the number of UGC sentences that would be comparable to those contained in a typical VOC study. They judged the human effort involved in extracting customer needs from 8,000 UGC sentences would be comparable, but slightly less than, the effort involved in extracting customer needs from interview transcripts.

The analysts, who extracted needs from the UGC, were drawn from the same marketing consulting firm that produced Figure 3. This enabled us to maintain a common level of training and experience. For each sentence, the analysts identified all customer needs in the sentence and coded those customer needs against the primary, secondary, and tertiary customer needs in the gold standard. If a tertiary customer need was not in the gold standard, the analysts attempted to assign the customer need to an existing secondary-customer-need group. If the tertiary customer need could not be assigned to a pre-existing customer-need group, the tertiary customer need was given a new number. This data set is unique because the analysts coded <u>all</u> customer needs in every sentence of the UGC. Typical practice does not maintain such a map between the source of each customer need and the customer need.

INFORMATION CONTAINED IN UGC VERSUS EXPERIENTIAL INTERVIEWS. We compared the information contained in the two sources of customer needs. This comparison is summarized in Figure 5a. Of the 86 tertiary customer needs extracted by human effort applied to the transcripts, 74 customer needs (86%) were extracted by human effort from the UGC. Importantly, analysts extracted seven new customer needs from the UGC, customer needs that were not extracted from the experiential interviews. This is impressive. We then asked analysts to examine an additional 4,000 randomly-selected UGC sentences to see if the customer needs, that were identified from experiential interviews, could be identified from additional UGC. Nine of the remaining twelve needs were identified. See Figure 5b. The analysts' supplementary task was limited; we do not know if the additional 4,000 sentences contained any additional customer needs. (We plan future research to identify the relative importances of the various customer needs.)

Figure 5. Comparison of Customer-Need Extraction from a Sample of UGC versus Experiential-Interview Transcripts

(a) Holding Extraction Costs for UGC to be Less than those for Experiential Interviews.

(b) Allowing Higher Extraction Costs for UGC, but Still Saving Interviewing Costs

We conclude that UGC is at least a comparable source of customer needs as experiential interviews. Because UGC eliminates the substantial effort cost involved in scheduling and implementing qualitative interviews, even with the additional 4,000 sentences, the total human-effort cost is less with the machine-human hybrid approach than with the human-only approach. We'll see later in this paper that machine-learning methods make extracting customer needs more efficient, thus enabling analysts to process a UGC corpus larger than 8,000 sentences for the same effort as was used to process transcripts. Further improvement should increase efficiency even more.

Human-effort coding of the 8,000-sentence UGC corpus suggests that 52% of the UGC sentences are informative about customer needs (contain an identified customer need). There was also high redundancy. Ten percent (10%) of the most-frequently mentioned customer needs were articulated in 54% of the informative sentences. These percentages suggest potential efficiency gains due to the CNN and clustering sentence representations.

CNN. When the training sample, X, is larger, the CNN can classify sentences better. Figure 6 plots the ability of the CNN to classify sentences as a function of X. Figure 6 reports results up to 6,000 sentences because preprocessing eliminated 1,394 sentences as too short or too long. This left 6,606 sentences eligible for use in training the CNN.

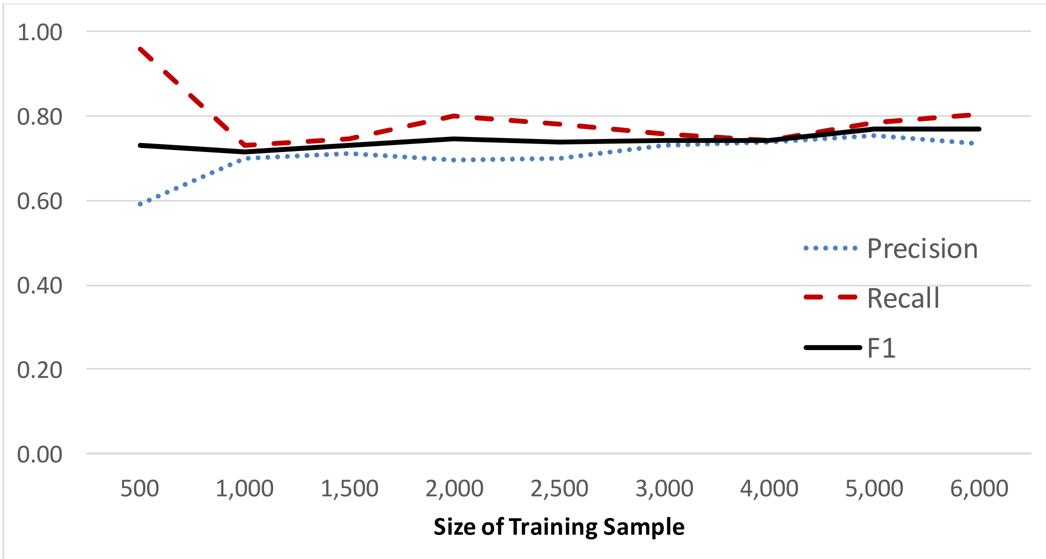
We report three statistics that are common in machine learning. Precision, in machine learning, is comparable to hit rates in conjoint analysis (and not to be confused with the scale factor in conjoint analysis). In sentence classification, precision is the percent of sentences that are informative given that they have been labeled as informative. Recall is the percent of informative sentences that were correctly labeled as informative. F_1 is a composite measure equal to:

$$F_1 = \frac{precision \cdot recall}{\frac{1}{2}(precision + recall)}$$

There are tradeoffs in precision and recall as the size of the training sample increases, but their impact on the composite measure, F_1, appears to stabilize around $X = 1{,}000$. At $X = 1{,}000$, Figure 6 reports a precision of 70% and a recall of 73%.

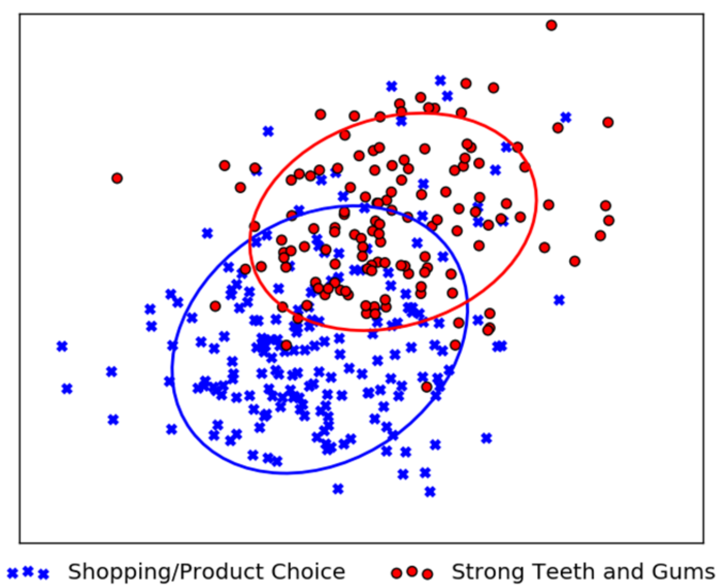
The CNN is effective if it identifies customer needs in the UGC corpus that were not in the training data. This was the case. The CNN identified customer needs in the UGC corpus that were not in the training data.

Figure 6. Precision, Recall, and F_1 as a Function of the Size of the Training Sample



CLUSTERS OF SENTENCE REPRESENTATIONS. To visualize whether or not clustering sentence representations enhanced diversity in customer needs, we use principle components analysis to project the sentence representations onto two dimensions. Information is lost, but we can see visually whether or not customer needs were separated by clustering sentence representations. Figure 7 reports the results.

Figure 7. Two-dimensional Projection of 300-Dimensional Sentence Representations



The red dots are sentence representations that were coded (by human judges) as belonging to the primary customer need of “strong teeth and gums.” The blue dots are sentence representations that were coded as “shopping/product choice.” The ovals represent the smallest ellipsis inscribing 90% of the corresponding set. Figure 7 suggests that, while not perfect, the clusters of sentence representations did achieve separation among customer needs.

GAINS IN EFFICIENCY DUE TO THE MACHINE-HUMAN HYBRID METHOD. We use our database to compare counterfactual simulations of the number of customer needs that would have been identified by various methods. We compare the methods for various numbers of sampled UGC sentences. We chose to train the CNN on 5,000 sentences to approximate how we expect the CNN to be used in practice. We believe the larger training sample eliminates randomness in our analysis, but we do not believe that the relative comparisons of methods would change.

When we train the CNN on 5,000 sentences, we can hold out 1,606 sentences after preprocessing to eliminate sentences that are too short or too long. At this $X = 5{,}000$, the CNN achieves a precision of 76%, a recall of 78%, and an F_1 of 77%. The CNN identifies 1,040 of the 1,606 sentences as informative.

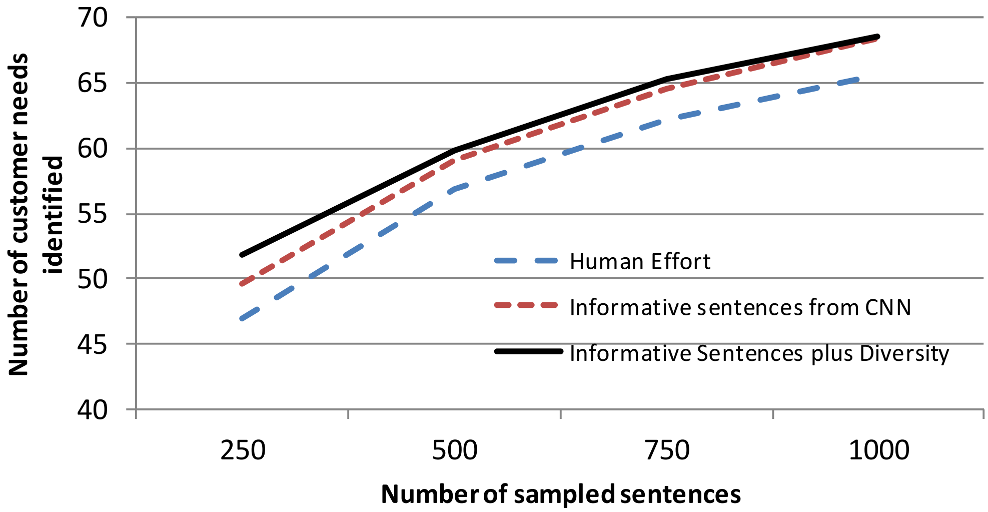
For each of three methods, we compute counterfactuals assuming the analysts have only the resources to review Y sentences for $Y = 250, 500, 750,$ and $1{,}000$. To compare to a human-effort benchmark, we evaluate the customer needs identified from a random selection from the UGC corpus (assuming preprocessing to eliminate sentences that are too short or too long). For example, an analyst would randomly select 250 sentences from the preprocessed corpus and review all 250 sentences. We redraw random samples 1,000 times and average. The results of random selection are shown in Figure 8 by a dashed blue line.

We improve efficiency by using the CNN to identify informative sentences. To test efficiency gains, we randomly select from informative sentences (dotted red line in Figure 8). We increase efficiency further by using the CNN to screen for informative sentences, clustering sentence representations, and selecting from sentence representations proportional to the size of the clusters (solid black line in Figure 8).

Over the range of the counterfactual simulations, Figure 8 suggests that the machine-learning stages enhance efficiency. There are gains due to using the CNN to eliminate non-informative sentences and additional gains due to using sentence representations to seek diversity within the corpus. The gains to diversity decrease with Y, but the gains due to the identification of informative sentences continue throughout the range of the counterfactual simulations.

We also interpret Figure 8 horizontally. Human effort requires, on average, 1,000 sentences to identify 65.6 customer needs. If we prescreen with machine learning to select diverse, informative sentences, an analyst can identify, on average, 65.2 customer needs from 750 sentences. These efficiencies represent a human-effort saving of 25%. Given that human-effort-based reviewing of experiential interviews has been optimized over almost thirty years of continuous improvement, these proof-of-concept results are promising. We expect the machine-learning methods, themselves, to be subject to continuous improvement as analysts learn, by trial and error, how best to merge machine learning with human effort.

Figure 8. Comparison of Efficiencies among Various Means to Select UGC Sentences for Review



DISCUSSION AND SUMMARY

A high-craft conjoint analysis study requires that attributes and attribute levels be chosen carefully. VOC methods are a proven method by which to identify a complete set of attributes. VOC methods identify customer needs, then established methods, such as QFD, hedonic regression, or the Brunswik lens model, select attributes that are solutions to customer needs.

In this paper we establish that machine-learning methods show promise to extract customer needs more effectively and more efficiently. Machine-learning methods also extract new customer needs that are missed by traditional experiential-interview studies. Once perfected, machine-learning methods applied to UGC will enable conjoint-analysis analysts to extract a more-complete set of customer needs (attributes) and do so quicker and with less human-effort costs.

UGC. Our results suggest that UGC can substitute for experiential interviews. In a limited corpus of 8,000 sentences, human analysts were able to extract roughly as many customer needs as would have been extracted from experiential interviews. The overlap was not perfect, but the UGC did identify customer needs not in the interview transcripts. A comparison of Figures 5a and 5b suggests that, with a larger corpus, particularly with efficiencies due to machine learning, UGC should provide sufficient information with which to extract a more-complete set of customer needs than the typical experiential-interview study.

CNN. The CNN successfully identified non-informative sentences. Future research might optimize the CNN.

SENTENCE REPRESENTATIONS. Clustering sentence representations increases diversity, especially for small samples. However, as the size of the sample of sentences to review increases, the machine-human hybrid gets close to an exhaustive set of needs and the value of diversity decreases.

EFFICIENCY GAINS. Perhaps the largest efficiency gain is the enhanced ability to replace experiential interviews with UGC. Experiential interviews are costly and require calendar time to recruit, schedule, and implement experiential interviews. A typical experiential-interview study requires about 4-5 weeks. UGC can be harvested quickly (less than a day) and at substantially lower cost.

We asked the marketing consulting firm to review 8,000 UGC sentences in depth because they judged that reviewing 8,000 UGC sentences was a conservative estimate of the effort required to review a typical set of experiential interviews. Even with 12,000 UGC sentences, the human effort for extraction is less than the human-effort in an experiential-interview study. Both the CNN and clustering sentence representations makes the review of the UGC sentences more efficient by as much as 25%. (A percentage we hope to increase with continuous improvement.)

MACHINE-LEARNING APPLIED TO INTERVIEW TRANSCRIPTS. There is nothing to prevent using the CNN and the sentence representation clusters on interview transcripts. We expect to see efficiencies there as well. The machine-human hybrid method applied to the interview transcripts can be useful in a product categories where UGC is either not available or not extensive.

SUMMARY. Understanding customer needs helps define a more-complete set of attributes and improve the quality of the conjoint study. Based on our initial proof-of-concept application, we are optimistic about the potential of UGC and machine learning to transform the practice of identifying customer needs. We feel that the CNN and sentence representations are uniquely suited to the analysis of UGC because these methods do more than count words. They look to deep semantic structure as is required in the analysis of UGC.

REFERENCES

Allenby, Greg M., Jeff Brazell, John R. Howell, and Peter E. Rossi (2014), “Valuation of Patented Product Features.” *Journal of Law and Economics*, 57:3 (August). 629-663.

Alsup, William (2012), “Order Granting in Part and Denying in Part Google’s Daubert Motion to Exclude Dr. Cockburn’s Third Report,” Oracle America, Inc. v. Google, Inc. No. C-10-03561 WHA, United States District Court for the Northern District of California, March 13.

Blei, David M, Andrew Y. Ng, and Michael I. Jordan (2003), “Latent Dirichlet Allocation,” *Journal of Machine Learning Research*, 3, 993-1022.

Brunswik, Egon (1952), *The Conceptual Framework of Psychology*, (Chicago, IL: Chicago University Press).

Cameron, Lisa, Michael Cragg, and Daniel McFadden (2013), “The Role of Conjoint Surveys in Reasonable Royalty Cases,” *Law360*, October 16.

Collobert, Ronan, Jason Weston, Leon Botto, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa (2011), “Natural Language Processing (Almost) From Scratch,” *Journal of Machine Learning Research*, 12, (Aug), 2493-2537.

Continuum (2016). https://www.continuuminnovation.com/en/what-we-do/case-studies/swiffer/.

Eggers, Felix, John R. Hauser, Matthew Selove (2016), “The Effects of Incentive Alignment, Realistic Images, Video Instructions, And Ceteris Paribus Instructions on Willingness to Pay and Price Equilibria,” *Proceedings of the 19th Sawtooth Software Conference*, Park City, UT, September 26-20.

Griffin, Abbie (1992), “Evaluating QFD's Use in US Firms as a Process For Developing Products,” *Journal of Product Innovation Management*, 9 (3), (September), 171-187.

Griffin, Abbie and John R. Hauser (1993), "The Voice of the Customer," *Marketing Science*, 12, 1, (Winter), 1-27.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum (2004), “Integrating Topics and Syntax,” *Advances In Neural Information Processing Systems*, 17, 537–544.

Hauser, John R. and Don Clausing (1988), "The House of Quality," *Harvard Business Review*, 66, 3, (May-June), 63-73.

Herrmann, Andreas., Frank Huber, and Christine Braunstein (2000), “Market-Driven Product And Service Design: Bridging The Gap Between Customer Needs, Quality Management, And Customer Satisfaction.” *International Journal of Production Economics*, 66 (1), 77-96.

Iyyer, Mohit, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III (2015), “Deep Unordered Composition Rivals Syntactic Methods For Text Classification,” *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (Volume 1: Long Papers), 1681-1691. (Beijing, China: Association for Computational Linguistics).

Kao Group (2016). http://www.company-histories.com/Kao-Corporation-Company-History.html.

Kim, Yoon (2014), “Convolutional Neural Networks For Sentence Classification,” arXiv:1408.5882v2 [cs.CL], Sept 3.

Kiss, Tibor and Jan Strunk (2006), “Unsupervised Multilingual Sentence Boundary Detection,” *Computational Linguistics*, 32(4), 485-525.

Le, Quoe and Tomas Mikolov (2014), “Distributed Representations of Sentences and Documents,” *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, 32, 1188-1196.

Lee, Thomas Y. and Eric T. Bradlow (2011), “Automated Marketing Research Using Online Customer Reviews,” *Journal of Marketing Research*, 48(5), 881-894.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). Efficient Estimation Of Word Representations In Vector Space,” arXiv:1301.3781v3 [cs.CL]m Sept 7,1301.3781.

Mikolov, Tomas, Ilya Sutskever, Kai Che, Greg S. Corrado, and Jeffrey Dean (2013b), “Distributed Representations of Words and Phrases and their Compositionality,” *Advances In Neural Information Processing Systems*, 26, 3111–3119.

Sullivan, Lawrence P. (1986). “Quality Function Deployment,” *Quality Progress*, 19(6), 39-50.

Tai, Kai Sheng, Richard Socher, and Christopher D. Manning (2015), “Improved Semantic Representations From Tree-structured Long Short-Term Memory Networks,” arXiv:1503.00075v3 [cs.CL], May 30.

Timoshenko, Artem and John R. Hauser (2017), "Identifying Customer Needs from User Generated Content," (Cambridge, MA: MIT Sloan School of Management).

Ulrich Karl T. and Steven D. Eppinger (2016) *Product Design and Development, 6E*. (New York, NY: McGraw-Hill).