Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Microsummarization of Online Reviews: An Experimental Study

Rebecca Mason *
Google, Inc.
Cambridge, Massachusetts
ramason@google.com

Benjamin Gaska *
University of Arizona
Tucson, Arizona
bengaska@email.arizona.edu

Benjamin Van Durme
Johns Hopkins University
Baltimore, Maryland
vandurme@cs.jhu.edu

Pallavi Choudhury, Ted Hart, Bill Dolan, Kristina Toutanova, and **Margaret Mitchell**
Microsoft Research
Redmond, Washington
{pallavic,tedhar,billdol,kristout,memitc}@microsoft.com

Abstract

Mobile and location-based social media applications provide platforms for users to share brief opinions about products, venues, and services. These quickly typed opinions, or *micro-reviews*, are a valuable source of current sentiment on a wide variety of subjects. However, there is currently little research on how to mine this information to present it back to users in easily consumable way. In this paper, we introduce the task of *microsummarization*, which combines sentiment analysis, summarization, and entity recognition in order to surface key content to users. We explore unsupervised and supervised methods for this task, and find we can reliably extract relevant entities and the sentiment targeted towards them using crowd-sourced labels as supervision. In an end-to-end evaluation, we find our best-performing system is vastly preferred by judges over a traditional extractive summarization approach. This work motivates an entirely new approach to summarization, incorporating both sentiment analysis and item extraction for modernized, at-a-glance presentation of public opinion.

Introduction

The proliferation of short thoughts and reviews in social media and mobile-based communication provides an easy way for people to communicate their opinions to the rest of the world. For example, the mobile application FourSquare (www.foursquare.com) contains user-submitted reviews of venues, such as restaurants and other businesses. As a FourSquare user approaches a venue, FourSquare supplies the user's phone with review snippets from others who have visited the same venue. These *micro-reviews* consist of brief, somewhat spontaneous observations on others' experiences at that venue (see Table 1). These brief texts are a rich source for determining consensus opinions about different items and places.

In this work, we analyze FourSquare reviews to create a short list for at-a-glance information about a venue. We explore several approaches to identify key items, determine the sentiment expressed towards these items, and summarize this content to a user. The proposed task, which we will refer to as *microsummarization*, therefore crosscuts the tasks of entity recognition, sentiment analysis, and summarization.

This research follows an "end-to-end" approach, and is complementary to research on specific subtasks within the end-to-end system: We are motivated by summarization for real users, as compared to existing efforts that drill down on a predefined subtask without realistically addressing how the task may work in practice. We therefore compare both state-of-the-art and novel methods in order to explore their utility in an applied end-to-end scenario, and include human evaluation of the final output. This work motivates a new view of summarization, focused on identifying key items and the public opinion about them.

Our approach begins with unsupervised clustering of reviews to automatically discover words used in similar contexts. From this, we define a set of *facets*[1] to identify in the reviews, and compare models for facet recognition that require different amounts of supervision.

This content is then tagged using a neural-network based sentiment model, and the sentiment and entities are used to create a summary for an end user. We provide automatic evaluation of each component compared against gold standards, as well as a crowdsourced study on the overall quality of the final microsummaries. We find our proposed system utilizing ClusterSum is preferred by users 80% of the time over traditional extractive summarization techniques.

This research makes the following contributions:

1. Introduces a new NLP task focused on presenting short, to-the-point summaries based on analysis of social online review text
2. Provides new approaches for summarization, ClusterSum

Try the Beach Salad!
My favorite is the Chicken Club. It is soooo goooood
Sad to hear my homie Sara doeat work here anymore :-(
They have #The12thCan!!
Urinals are perfect if you're a midget

Table 1: Examples of micro-reviews from FourSquare.

*Work done during internship at Microsoft Research.

[1] Also known as *entities*, *attributes*, or *aspects*.

and FacetSum, tailored to the micro-review domain

3. Presents an end-to-end system incorporating sentiment analysis, summarization, and facet recognition
4. Evaluates the utility and quality of the microsummaries using crowdsourcing
5. Provides the first empirical comparison between spectral two-step CCA embeddings and word2vec embeddings on a facet (entity) recognition task.

In the next section, we discuss related work. We then discuss our approaches to summarization, facet recognition, and sentiment analysis. Finally, we present a crowdsourced evaluation of the microsummarization end task. We find that an end-to-end system incorporating a semi-CRF with neural embeddings and a neural-network based sentiment module works well for this task, and a crowdsourced study suggests users would enjoy the proposed technology.

Related Work

The current task is most closely related to the task of *aspect-based sentiment summarization* (Titov and McDonald 2008; Gamon et al. 2005), which takes as input a set of user reviews for a predefined entity, such as a product or service, and produces a set of relevant aspects of that entity, the aggregated sentiment for each aspect, and supporting textual evidence. The current work expands from this previous work significantly: We assume nothing about the format or rating structure of the mined user reviews. We propose an end-to-end system that is *open-domain*, without being limited to, e.g., a small set of predefined entities. The current work extends to include both the initial definition of entities as well as the final presentation to users. Further, in addition to well-established automatic metrics to evaluate each component, the end-to-end task is evaluated with potential users. The methodology proposed in this work is shown to outperform alternative approaches.

Summarization While there is much previous work on summarizing micro-blogs such as Twitter (O'Connor, Krieger, and Ahn 2010; Chakrabarti and Punera 2011; Liu, Liu, and Weng 2011), the focus is typically on understanding the source content, and not on extracting key items to surface to an end user. Recent work from (Nguyen, Lauw, and Tsaparas 2015) demonstrates a method for synthesis of full reviews from collections of micro-reviews. Their focus and presentation differ significantly from ours in that they create full review documents, without any attempt to identify and display pertinent items from the micro-reviews.

A number of approaches to review summarization use probabilistic topic models to capture the facets and sentiments expressed in user-written reviews, leveraging further user annotations. In contrast to previous work, we explore approaches that do not rely on additional information to be supplied by users writing the reviews. We argue that placing minimal requirements on user-provided supervision is critical in work that seeks to summarize the content of the *quickly typed opinions* that characterize mobile and social online reviews. (Titov and McDonald 2008) and (Mcauliffe and Blei 2008) incorporate some supervision in the form of scores for each aspect of the entity being reviewed, and (Branavan et al. 2009) jointly model review text and user-defined keyphrases.

Further, it is worth discussion that the brevity of micro-reviews affects the quality of the topics that can be induced by topics models such as latent Dirichlet allocation (LDA), which are commonly used for summarization of longer documents. LDA works well to pick out thematically related items within a topic, which has the net effect of distinguishing more strongly between types of venues (e.g., Chinese cuisine vs. Mexican cuisine), rather than types within a venue (e.g., service vs. amenities). Brody and Elhadad (2010) attempted to overcome this issue by using a local (sentence-based) LDA model. Their work was not available for comparison, however, when we implemented our own version of their model,[2] we found that depending on the number of topics, the topics were either incoherent or grouped words together that we aimed to separate, e.g., service and food words.

Brody and Elhadad (2010) further address this issue by applying complex post-processing to pull out "representative words" for each topic. As we will further detail below, we are able to extract coherent groups of words in one pass by using Brown clustering (Brown et al. 1992). Examples of word groups selected by LDA and Brown clustering are shown in Table 2. Both LDA and Brown clustering can be used to group related words without supervision, however, the two methods model very different views of the data. We find the view of the data provided by Brown clustering to be compelling for this task.

Named Entity Recognition In this work, we develop a model to identify and classify the relevant facets of micro-reviews to incorporate into the summary. The task of named entity recognition (NER) is therefore also relevant, and indeed, NER in micro-blogs – particularly Twitter – has gathered attention in recent work (Ritter et al. 2011; Derczynski et al. 2013). Our work is most similar to (Liu et al. 2011), which combines a k-nearest neighbor (KNN) approach with a conditional random field (CRF)-based labeler to combat the sparsity and noisiness of micro-blog text. However, such work builds off previous named entity work by using, e.g., predefined entity types and external gazetteers that utilize such types. In contrast, given the goal of discovering what people talk about in reviews, *whatever those things may be*, the current work proposes a robust method for defining novel facets. Further, due to the relative novelty of this task, publicly available curated gazetteers do not exist; the current work is necessary to help build them.

Word Representations As we will further detail, we explore the use of unsupervised word features such as Brown clusters (Brown et al. 1992) to generalize a small amount of annotated training data. Recently there has been substantial work on integrating discrete word clusters and continuous word representations as features for entity recognition and other tasks (Turian, Ratinov, and Bengio 2010). Spectral two-step CCA (Dhillon et al. 2012) embeddings and word2vec embeddings (Mikolov et al. 2013) have both been shown to

[2] And further development, including the use of an asymmetric prior to encourage more coherent topics.

(a)

Topic	Top Words
2	sushi roll tuna spicy fresh tempura sashimi sake japanese
5	coffee latte iced mocha chocolate chai espresso hot cup tea cafe pumpkin
6	cream ice chocolate cake butter peanut yogurt cheesecake flavors cookies red cookie cupcakes vanilla
11	tacos burrito taco mexican salsa chips margaritas margarita guacamole burritos chicken queso fish
22	tea sweet strawberry green lemonade mango smoothie

(b)

Cluster	Top Words
000011100110	margarita martini smoothie mojito mimosa slush lassi marg daiquiri slushie gimlet slushy cosmo Cosmo caipirinha fizz
001011000000	sandwiches burritos sandwhiches pitas skillets
0001010010111	vanilla caramel mint hazelnut peppermint Carmel carmel matcha eggnog cardamom kona caffe chia zebra
00010100110	chocolate fudge choc cocoa java coco
000101001110	butter Butter crumb brittle buster
000101001111	chip mousse fortune chunk Chip mouse ganache coca Coca

Table 2: Example of top words in relevant (a) LDA Topics vs. (b) Brown clusters. LDA tends to group thematically related words together, e.g., "margarita" and "guacamole" may belong to a "Mexican restaurant" category; Brown clustering tends to group syntactically/semantically related words together, e.g., "margarita" and "smoothie" may belong to a "drinks" category.

outperform previously proposed Collobert and Weston (2011) embeddings (Pennington, Socher, and Manning 2014). We therefore utilize both word2vec and two-step CCA embeddings. To the best of our knowledge, this work provides the first empirical comparison between these two methods on a recognition task.

Problem Formulation

In this paper, we explore the problem of micro-review summarization in the domain of restaurant reviews. We define our task as follows: Given a set of micro-reviews about some restaurant venue, extract *positive items* (things to try, people to bring, etc.) and *negative items* (things to avoid, problems to be aware of, etc.), organized by *facet type* (venue, food, etc.). Our data consist of micro-reviews for which the restaurant being reviewed has already been identified.

For our experiments, we collect data from FourSquare.com. FourSquare is a mobile application where users "check in" at various locations, and leave "tips" for other users consisting of recommendations or things to try. We collect information from 818,701 venues that have been labeled by FourSquare users as food or drink establishments in the United States, which total 6,073,820 "tips" (micro-reviews). The average length of a tip is 14.78 tokens, including punctuation and emoticons. Tips are short and to the point, but are sometimes irrelevant, inappropriate, or contain irregular spelling or vocabulary due to the affordances of typing on a mobile phone.

Facet Recognition

This section presents methods for defining facets and recognizing facet entities mentioned in micro-reviews. We compare a baseline method using unsupervised Brown clustering and a set of heuristics to a supervised Semi-CRF method that incorporates unsupervised word features.

Defining the Facets

The start of the microsummarization task begins with defining the facet types. We begin with unsupervised Brown clustering (Brown et al. 1992), which has been used to good effect for NLP tasks such as named entity recognition (Miller, Guinness, and Zamanian 2004) and dependency parsing (Koo, Carreras, and Collins 2008; Spitkovsky et al. 2011). Brown clustering is a greedy hierarchical algorithm that finds a clustering of words that maximizes the mutual information between adjacent clusters – in effect, learning a class-based bigram language model. Each word type is assigned to a fine-grained cluster at a leaf of the hierarchy of clusters. Each cluster can be uniquely identified by the path from the root cluster to that leaf. Representing this path as a bit-string (1 indicating left, 0 indicating right) allows a simple coarsening of the clusters by truncating the bit-strings.

Facet	Example
Amenities	*trivia night, patio seating, free wifi ,clean bathrooms*
Customers	*couples, the kids, coworkers*
Events	*anniversaries, a blind date, the bachelorette party*
Food	*a bacon cheeseburger, sushi, the craft beer selection*
Service	*the bartenders, baristas, our server Tom*
Venue	*an intimate atmosphere, a French cafe*

Table 3: Examples of items for restaurant facets.

del, de, one, all, some, any, each, of, a, an, the, this these, those, that, my, her, his, their, our, your, other, only

Table 4: Function words permitted in unsupervised facet recognition with heuristics.

We trained 1000 Brown clusters, and had two native English speakers mark possible categories for each. We define a set of *facets* from the set of labels where both annotators agreed. Total annotation time for this task was under 4 hours. Table 3 shows examples of entities for each of the facets.[3]

Unsupervised Model

For a baseline model, we leverage the unsupervised Brown cluster labels given by the annotators and a set of heuristics to label the reviews. Borrowing from the B-I-O labeling scheme used in named entity recognition, we label each word as Outside the facet span, Inside, or Beginning the span.

Facet recognition in this method follows a set of simple heuristics. Each review is scanned right-to-left: When a word is discovered that belongs to one of the labeled Brown clusters, that word is marked I and labeled according to the cluster.

[3] These facet labels – *Amenities, Customers, Events, Food (and Drink), Service,* and *Venue* – roughly correspond to the facets found by Brody and Elhadad (2010) after post-processing.

Facet	Amenities	Customers	Events	Food	Service	Venue
Consensus	63.7	71.4	44.7	96.6	92.2	81.2

Table 5: Percentage of spans marked and given the same label by two or more workers.

Each preceding word that is in a cluster with the same label also receives that label.[4] We include a small set of function words as part of the span, listed in Table 4. The last word right-to-left that is in a cluster or the permitted function words is marked as B for the given facet, and the search for the next item continues.

Supervised Model

Semi-Markov CRF, or 'semi-CRF' (Sarawagi and Cohen 2004), is a type of conditional random field (Lafferty, McCallum, and Pereira 2001) in which labels are assigned to subsequences of the input sequence, rather than to individual elements. This formulation allows features to be defined for sequences of consecutive word tokens, and allows for the boundaries of arbitrarily long expressions to be modeled.

Under the usual first-order Markovian assumption, the feature functions $g_i(j, \mathbf{x}, \mathbf{s})$ are functions of an observed sentence $\mathbf{x}$, the current segment s_j, and the label of the previous segment. The conditional probability of a segmentation $\mathbf{s}$ given the sentence $\mathbf{x}$ and the feature functions g_i with corresponding weight λ_i is defined as:

$$p(s|x) = \frac{1}{Z(\mathbf{x})} \exp\left\{ \sum_j \sum_i \lambda_i g_i(j, \mathbf{x}, \mathbf{s}) \right\} \quad (1)$$

A segmentation $\mathbf{s}$ is a sequence of facet segments and non-facet segments, with non-facet segments represented as unit-length segments tagged as O.

Annotation of Named Entities To collect training data for the Semi-CRF model, we developed a crowdsourced annotation task to identify entities in micro-reviews. Workers were instructed to look at the micro-reviews and highlight items for each of the six facets (see Table 3). A screenshot of the user interface is included with the supplemental material. 13,000 micro-reviews were annotated by 3 workers each. From this, we extract the maximally overlapping spans given the same label by at least 2 annotators, yielding 10,712 annotated reviews. The percentage of times a span labelled by one worker is labelled by two or all three workers is shown in Table 5. Workers have highest consensus for Food & Drink (96.56%) and Service (92.16%) entities. Overall, about 75% of the micro-reviews had the same spans marked by at least two annotators.

We implement the following features:

Unigram Features We include words and their lowercase forms for all segment words except for infrequent words (seen 10 times or less), which we replace with indicator features following the approach in (Petrov et al. 2006). We also specially mark which are the first, second, second-to-last, and last words of the segment.

Segment Features An advantage of the Semi-CRF model is that features can be defined for sequences of words, rather than just for each word individually. These features include the number of words in the sequence, capitalization and punctuation patterns for the entire sequence, and whether the sequence contains punctuation. We also include character n-grams of length 3-5.

Unsupervised Word Features Finally, we include vector word embeddings computed using two different techniques, word2vec (Mikolov et al. 2013) and a spectral-based encoding, Two-Step CCA (Dhillon et al. 2012). Each of these unsupervised word representations are trained on 5,772,600 micro-reviews. Testing on a held-out set suggested that optimal settings for word2vec were 40 dimensions and the continuous bag-of-words (CBOW) algorithm with a window size of 1. We also include features based on the Brown clusters, using various prefix lengths for the bitstrings (4, 8, 12, and all).

Results: Facet Recognition

Results are shown in Table 6. We find that a word2vec representation significantly outperforms spectral embeddings on the most frequent food facet.[5] However, spectral embeddings have fewer false positives for the less frequent facets. We can also see that all variants of the supervised Semi-CRF model that use cost-effective annotation of training data via crowdsourcing achieve substantially higher performance than the unsupervised method. The unsupervised features (Brown, Spectral, and word2vec) offer large improvements over a basic model with unigram and segment features. Interestingly, recall is much higher across the board at the word level without unsupervised features.

Review Summarization

A key question in microsummarization is how to identify relevant content. We seek to maximize the likelihood of each entity we select with respect to all of the reviews for the same venue. Identifying these entities is difficult due to noise and variation in how they are described. This is especially a problem in the restaurant domain: official names of dishes – such "Japanese Snapper with Sea Urchin and Coconut Risotto" – are often not in alignment with references in micro-reviews. Additionally, there are reviews that our model should ignore, such as deceptive reviews and spam.

To combat this problem, we limit the set of micro-reviews that we extract entities from. Rather than performing facet recognition on every micro-review for a venue, instead we only extract entities from micro-reviews that are *most representative* of reviews for that venue. Research in extractive summarization focuses on solving precisely this problem, isolating sentences that best reflect the content of the whole document. We adapt such models to the review domain, extracting key reviews that best represent the language used by all reviewers of a venue. We compare two well-known

[4] We follow this approach in English because it is dominantly right-branching: The head word of a phrase is usually at its end.

[5] p-value < 0.05 according to a paired sign test for sentence-level f-measure.

PHRASE LEVEL

Approach	Food			Service			Venue			Amenities			Events			Customers		
	Pr.	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**
Unsupervised+Heur	0.48	0.51	0.49	0.55	0.44	0.49	0.30	0.12	0.17	0.11	0.13	0.12	0.07	0.14	0.09	0.35	0.41	0.38
Semi-CRF (Basic)	0.72	0.67	0.69	0.59	0.60	0.59	0.15	0.19	0.17	0.10	0.11	0.10	0.00	0.00	0.00	0.03	0.06	0.04
Semi-CRF w/ Brown	0.77	0.76	0.77	0.78	0.72	0.75	0.56	0.52	0.54	0.69	0.39	0.50	0.00	0.00	0.00	0.75	0.18	0.29
Semi-CRF w/ Brown+Spectral	0.79	0.76	0.77	0.78	0.70	0.74	0.58	0.49	0.53	0.73	0.35	0.47	0.00	0.00	0.00	0.80	0.24	0.37
Semi-CRF w/ Brown+Word2Vec	0.82	0.77	0.79	0.78	0.71	0.74	0.60	0.48	0.53	0.65	0.35	0.46	0.00	0.00	0.00	0.57	0.24	0.34

WORD LEVEL

Approach	Food			Service			Venue			Amenities			Events			Customers		
	Pr.	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**	**Pr.**	**Rec.**	**F1**
Unsupervised+Heur	0.83	0.65	0.73	0.85	0.59	0.70	0.55	0.15	0.24	0.24	0.18	0.20	0.11	0.13	0.12	0.36	0.33	0.34
Semi-CRF (Basic)	0.79	1.00	0.88	0.62	1.00	0.77	0.36	0.97	0.52	0.38	0.85	0.53	0.62	0.53	0.57	0.17	0.97	0.29
Semi-CRF w/ Brown	0.86	0.92	0.89	0.87	0.78	0.82	0.64	0.66	0.65	0.77	0.51	0.61	0.33	0.07	0.11	0.75	0.20	0.32
Semi-CRF w/ Brown+Spectral	0.88	0.92	0.90	0.88	0.77	0.82	0.67	0.62	0.64	0.86	0.50	0.63	0.33	0.07	0.12	0.08	0.23	0.12
Semi-CRF w/ Brown+Word2Vec	0.92	0.90	0.91	0.88	0.76	0.82	0.71	0.62	0.66	0.80	0.54	0.65	0.33	0.07	0.12	0.58	0.23	0.32

Table 6: Precision, Recall, and F-score for facet recognition. We compare our unsupervised model with manual heuristics (Unsupervised+Heur) to several semi-CRF models. The semi-CRF model incorporating word2vec features perform well, however, the spectral method had less false positives for facets with few instances.

summarization methods, SumBasic and KLSum, and introduce two further methods, which we call *ClusterSum* and FacetSum, for item-specific summarization.

SumBasic SumBasic (Nenkova and Vanderwende 2005) is an algorithm for extractive multi-document summarization. It generates a summary by selecting sentences from the source documents, with the exclusive objective of maximizing the appearance of non-function words that have high frequency in the source documents.

For each micro-review, a SumBasic score is computed based on word w frequency. Using r to represent each micro-review, and $\mathcal{V}$ for the set of micro-reviews for a venue:

$$Score(r) = \sum_{w \in r} \frac{1}{|r|} p(w|\mathcal{V}) \qquad (2)$$

We select the top six micro-reviews for each venue, and to encourage diversity, include a decay function for the probability of words after each selection.

KLSum We implement the KLSum method following the definition in (Haghighi and Vanderwende 2009), with values of $p(w|\mathcal{V})$ smoothed by .01 to ensure there are no zeros in the denominator. The best summary is selected using a greedy search. As with SumBasic, we select six micro-reviews are selected for each venue.

ClusterSum Clustering is used to increase the diversity of content in extractive multi-document summaries (Otterbacher, Erkan, and Radev 2005; Qazvinian and Radev 2008). Clustering algorithms group together text with similar content, where larger clusters represent more meaningful content. Redundancy can be avoided by selecting text from different clusters. For this summarization approach, we cluster microreviews with similar content, then select two micro-reviews from each of the three largest clusters. [6]

As a similarity measure for micro-reviews, we use cosine similarity of non function-words. However, due to the brevity and sparse vocabulary of the micro-reviews, cosine similarity of vocabulary counts is not a sufficient metric of content similarity. Therefore, we employ an additional clustering on the words themselves.

We present in this paper ClusterSum, a two-stage clustering technique for aggregation of microreviews for summarization. The first stage of our ClusterSum approach replaces the words in each micro-review with its unsupervised word representation, to reduce the dimensionality of the data. The Brown cluster model is trained on 5772600 micro-reviews, with 1000 clusters, and the bitstrings for each cluster are truncated to 12 bits. In the second stage, we represent each tip for the venue as a vector of its word representations, and employ k-means clustering (k=5) on the micro-reviews for each venue. The KLSum method is then used to select top reviews from the clusters.

FacetSum Finally, we consider a novel means of summarization created for this paper as an alternative to traditional summarization techniques. Instead of summarizing over the entirety of the documents, FacetSum works directly on the extracted entities for *every* micro-review for a venue. Then we apply the clustering=based summarization approach discussed above, using extracted entities in place of documents.

Sentiment Analysis

Our method for determining the sentiment targeted at extracted entities is based on an adaptation of the Coooolll system (Tang et al. 2014), one of the top performing systems in the SemEval 2014 Twitter sentiment classification task. This system uses a neural network to analyze sentiment at the sentence level. We are interested in more fine-grained sentiment analysis, determining sentiment targeted at a specific entity rather than classifying the overall sentiment of a sentence. This allows us to tease out the different sentiment expressed in a single sentence. For example, in a sentence such as "the staff was awesome, but the fries were soggy", the sentiment towards *staff* is positive, but the sentiment towards *fries* is negative.

To adapt the sentiment system to this task, we explored ways to split each micro-review into phrases that contain one

[6] Fewer micro-reviews may be selected in some instances depending on the size of the clusters.

Experiment	Precision			Recall			F1			Accuracy
	Pos	Neg	Neu	Pos	Neg	Neu	Pos	Neg	Neu	All
Basic	89.2	78.4	**33.5**	83.6	64.5	54.1	**86.3**	70.8	41.4	77.8
NEU → POS	87.1	**89.0**	32.7	**85.4**	58.9	49.5	**86.3**	70.9	39.4	**77.9**
POS@1.5NEG, NEU@1.0NEG	**94.0**	65.4	31.6	73.9	**80.7**	**67.9**	82.8	**72.2**	**43.2**	74.1

Table 7: Precision, Recall, F1, and overall Accuracy % on sentiment classification of facets. Shown options are NEU → POS, Neutral changed to Positive for training; and POS@1.5NEG, NEU@1.0NEG, with Positive sampled as 1.5 x Neg and Neutral in equal amount to Neg. In low confidence cases ($< .0001$), the system suggests a Neutral label.

or more entities with the same sentiment targeted at them. The splitting is done at points where there are characters in the set {,, ;, :, ., !, ?} or words in the set {but, also, or} that are not within an entity. This simple approach is extremely effective, yielding phrases with the same sentiment expressed towards its entities for 86% of our training data.

We use crowdsourcing to collect sentiment labels, with a minimum of 3 unique judges, with up to 2 more added if there is not consensus. Judges were presented with the original review along with an extracted entity and asked whether the reviewer's attitude towards that entity is positive, negative, or neutral.[7] Splitting our training/testing data into phrases with entities with consensus targeted sentiment yielded 9617 sentiment phrases for training, 954 for testing.

The data collected from this task was naturally unbalanced, with 72% of phrases labeled as POSITIVE targeted sentiment, and only 15% labeled as NEGATIVE. We imagine that the sentiment is even more skewed than this distribution suggests, as a micro-review in isolation may appear neutral (e.g., a one-word review such as "BURGERS"); but in these cases, the fact that an entity is mentioned on a review site at all suggests that the reviewer felt positively towards it.

We therefore explored several methods to avoid bias in the data set towards positive mentions of entities and more definitively classify entity sentiments, downsampling the common entities and sampling different ratios of POSITIVE:NEGATIVE and NEUTRAL:NEGATIVE. We found that replacing all spans in the training data labeled NEUTRAL with a POSITIVE label increased our precision for NEGATIVE classification, and increased our recall for POSITIVE classification. Our most promising results as measured by F-score have POSITIVE training instances sampled at 1.5 times NEGATIVE training instances, with an equal amount of NEGATIVE and NEUTRAL training instances (see Table 7).

End-to-End: Evaluation

We construct an end-to-end system for each extractive summarization approach: SumBasic, KLSum, ClusterSum, and FacetSum. With each, the top reviews are extracted and fed into the best performing components for facet recognition and sentiment analysis. Descriptive statistics for the four approaches are shown in Table 9.

We use the Semi-CRF model with Brown clusters and word2vec features to extract facet items, and NEU→POS sentiment to determine the sentiment targeted at them. The facet items are presented for each venue as lists of *Positives* and *Negatives*, subdivided into each of the relevant facet types. Duplicate facet items for a given sentiment polarity are merged. This forms the final **microsummary**. Example output from an end-to-end system is shown in Figure 1.

Venue: Alessandro's
Positives:
Food:
Pasta
Veal marsala
Penne all vodka
Service:
The bartender Sam
Venue:
Parking
Negatives:
Amenities:
The bathroom

Figure 1: Example end-to-end sytem output

We use a crowdsourced study to compare the four summarization methods and determine user opinion. Trials were presented in randomized order, such that each judge could be exposed to each summarization method. Judges were asked which approach they preferred, and for each, if it were available in an app, would they use it. An example trial is shown in the supplemental materials. We randomly selected 125 venues, and in each trial, presented one microsummary and the corresponding extractive summary side-by-side in randomized order. Summaries were therefore matched such that the microsummary contained the facet items and sentiment identified in the corresponding extractive summary. As before, each task was presented to 3 unique judges, with up to 2 more judges added if there is not consensus. We evaluate user preference over all cases with majority agreement.

Results are shown in Table 8. Even with imperfect sentiment and facet recognition, we find that users strongly prefer microsummaries over extractive summaries using the same reviews. This is an interesting result, as it suggests a shift in the kinds of summaries that may be useful to explore moving forward. Specifically, this suggests a user preference for easy-to-read microsummaries, over traditional extractive summaries. Additionally, 49% of the judges reported that they would use the ClusterSum approach if it were available in an application (compared to 32% of judges who would use the corresponding extractive summarization approach). In comparison, for both KLSum and SumBasic, 41% of the judges said they would use this technology.

Significance testing between the systems suggests that the preference for microsummarization depends on the summarization approach (χ^2=13.96, p<.05), however, when limited to the top 3 systems (SumBasic, KLSum, and ClusterSum), the preference is independent of the approach (χ^2=2.5, p>.05); the preferences for microsummarization in these

[7] Judges were presented with full sentences, not the phrases extracted for training.

System	Preferred Approach			
	Micro	Extract	Same	No Consensus
SumBasic	78.4%	17.6%	0.8%	3.2%
KLSum	79.2%	14.4%	1.6%	4.8%
ClusterSum	80.0%	16.0%	0.0%	4.0%
FacetSum	68.0%	29.6%	0.0%	2.4%

Table 8: Consensus preference for each summarization approach. Judges chose between Microsummarization (Micro), Extractive Summarization (Extract), and "About the Same" (Same). % of restaurants without consensus from judges are reported in "No Consensus". The ClusterSum method introduced in this paper performs comparably to traditional summarization approaches.

System	SumBasic	KLSum	ClusterSum	FacetSum
Avg. # reviews	6.0	6.0	5.2	6.2
Avg. # entities	8.0	8.5	9.3	3.7

Table 9: Descriptive statistics on system output shown to users.

three systems are roughly equivalent.

Discussion

We have introduced a new microsummarization NLP task focused on presenting short, to-the-point summaries based on analysis of social online review text. Our end-to-end system for this task incorporates sentiment analysis, summarization, and facet recognition. We compared recent approaches to using word embeddings as features in a semi-markov CRF for facet recognition, and introduced novel methods for generating microsummaries. A final crowdsourced study showed that the microsummaries are strongly preferred over extractive summaries, and that a large fraction of users would use the technology if available.

References

Branavan, S.; Chen, H.; Eisenstein, J.; and Barzilay, R. 2009. Learning document-level semantic properties from free-text annotations. *Journal of Artificial Intelligence Research* 34(2):569.

Brody, S., and Elhadad, N. 2010. An unsupervised aspect-sentiment model for online reviews. In *NAACL*, 804–812. Association for Computational Linguistics.

Brown, P. F.; Desouza, P. V.; Mercer, R. L.; Pietra, V. J. D.; and Lai, J. C. 1992. Class-based n-gram models of natural language. *Computational linguistics* 18(4):467–479.

Chakrabarti, D., and Punera, K. 2011. Event summarization using tweets. In *ICWSM*.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537.

Derczynski, L.; Ritter, A.; Clark, S.; and Bontcheva, K. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *RANLP*, 198–206.

Dhillon, P.; Rodu, J.; Foster, D.; and Ungar, L. 2012. Two step cca: A new spectral method for estimating vector models of words. *arXiv preprint arXiv:1206.6403*.

Gamon, M.; Aue, A.; Corston-Oliver, S.; and Ringger, E. 2005. Pulse: Mining customer opinions from free text. *Proc. of the 6th Intl. Symposium on Intelligent Data Analysis*.

Haghighi, A., and Vanderwende, L. 2009. Exploring content models for multi-document summarization. In *ACL*, 362–370. Association for Computational Linguistics.

Koo, T.; Carreras, X.; and Collins, M. 2008. Simple semi-supervised dependency parsing. In *ACL*.

Lafferty, J.; McCallum, A.; and Pereira, F. C. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Liu, X.; Zhang, S.; Wei, F.; and Zhou, M. 2011. Recognizing named entities in tweets. In *ACL*, 359–367. Association for Computational Linguistics.

Liu, F.; Liu, Y.; and Weng, F. 2011. Why is sxsw trending?: exploring multiple text sources for twitter topic summarization. In *Proc. of the Workshop on Languages in Social Media*, 66–75. Association for Computational Linguistics.

Mcauliffe, J. D., and Blei, D. M. 2008. Supervised topic models. In *NIPS*, 121–128.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Miller, S.; Guinness, J.; and Zamanian, A. 2004. Name tagging with word clusters and discriminative training. In Dumais, S.; Marcu, D.; and Roukos, S., eds., *NAACL*.

Nenkova, A., and Vanderwende, L. 2005. The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

Nguyen, T.-S.; Lauw, H. W.; and Tsaparas, P. 2015. Review synthesis for micro-review summarization. *WSDM*.

O'Connor, B.; Krieger, M.; and Ahn, D. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Otterbacher, J.; Erkan, G.; and Radev, D. R. 2005. Using random walks for question-focused sentence retrieval. In *EMNLP*, 915–922. Association for Computational Linguistics.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. *EMNLP* 12.

Petrov, S.; Barrett, L.; Thibaux, R.; and Klein, D. 2006. Learning accurate, compact, and interpretable tree annotation. In *ACL*, 433–440. Association for Computational Linguistics.

Qazvinian, V., and Radev, D. R. 2008. Scientific paper summarization using citation summary networks. In *CoLing*, 689–696. Association for Computational Linguistics.

Ritter, A.; Clark, S.; Etzioni, O.; et al. 2011. Named entity recognition in tweets: an experimental study. In *EMNLP*, 1524–1534. Association for Computational Linguistics.

Sarawagi, S., and Cohen, W. W. 2004. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, 1185–1192.

Spitkovsky, V. I.; Alshawi, H.; Chang, A. X.; and Jurafsky, D. 2011. Unsupervised dependency parsing without gold part-of-speech tags. In *EMNLP*.

Tang, D.; Wei, F.; Yang, N.; Zhou, M.; Liu, T.; and Qin, B. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *ACL*, 1555–1565.

Titov, I., and McDonald, R. T. 2008. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, volume 8, 308–316. Citeseer.

Turian, J.; Ratinov, L.; and Bengio, Y. 2010. Word representations: a simple and general method for semi-supervised learning. In *ACL*, 384–394. Association for Computational Linguistics.