Extracting and Ranking Travel Tips from User-Generated Reviews

Ido Guy
Ben-Gurion University of the Negev
eBay Research
Israel*
idoguy@acm.org

Avihai Mejer
Yahoo Research
Israel
amejer@yahoo-inc.com

Alexander Nus
Yahoo Research
Israel
alexnus@yahoo-inc.com

Fiana Raiber
Technion – Israel Institute of Technology
Israel
fiana@tx.technion.ac.il

ABSTRACT

User-generated reviews are a key driving force behind some of the leading websites, such as Amazon, TripAdvisor, and Yelp. Yet, the proliferation of user reviews in such sites also poses an information overload challenge: many items, especially popular ones, have a large number of reviews, which cannot all be read by the user. In this work, we propose to extract short practical tips from user reviews. We focus on tips for travel attractions extracted from user reviews on TripAdvisor. Our method infers a list of templates from a small gold set of tips and applies them to user reviews to extract tip candidates. For each attraction, the associated candidates are then ranked according to their predicted usefulness. Evaluation based on labeling by professional annotators shows that our method produces high-quality tips, with good coverage of cities and attractions.

1. INTRODUCTION

User-generated reviews have become a popular medium for expressing opinions and sharing knowledge about items such as products (as in Amazon) and travel entities (as in TripAdvisor). Reviews have been shown to play a key role in users' decision making process and in the business success of reviewed items [6, 44, 45]. Yet, as is the case with other types of user-generate content (UGC), the success of reviews also leads to information overload. The vast amounts of user reviews accumulated for popular items, with each review usually containing multiple sentences, makes them practically impossible to consume. As a result, users often read only a few reviews and may miss helpful information. Current approaches to handle this issue range from sorting or ranking reviews by various criteria, such as date, number of helpful votes, or strength of the social tie to the reviewer [15, 23], through filtering by various parameters, such as user type, time of year, or extracted phrases or words [43], to applying summarization techniques of key features, opinions, or concepts [14, 37].

In this work, we propose to extract one-sentence tips from a large collection of reviews. We refer to a *tip* as a concise piece of practical non-obvious self-contained advice, which may often lead to an action [39][1]. Previous work has already indicated that users view the ability to receive tips and recommendations as one of the key benefits of UGC [3]. We argue that in certain scenarios, users may be more interested in such tips, rather than in the entire content of the review, which may include lengthy descriptions, historical facts, or personal experiences. Tips may especially come in handy for small-screen mobile device users, who are often short in time and may desire to get the gist of the crowd's word of advice about a place they are planning to visit or an item they want to buy.

Despite the potential value of tips, they are not as abundant as user reviews. One of very few examples of an application that has adopted the short tip notion and eschewed reviews is the location service Foursqaure, which allows its users to write short tips when they occur to a place. Yelp has introduced the notion of tips to its mobile application users, but these have not become as nearly as popular as reviews. TripAdvisor introduced tips as part of its city guides, but the coverage of these tips is low. While directly collecting tips from the community can be valuable, we believe that automatic extraction can help overcome the cold start problem [39].

Our work focuses on TripAdvisor, which is among the most popular sources of travel information [30, 42], incorporating over 385 million user reviews. When planning a trip, reviews on TripAdvisor are often perceived by users as more reliable, enjoyable, and up-to-date compared to other information sources [11]. While TripAdvisor contains reviews for a variety of travel entities, such as hotels, flights, and restaurants, we focus on tourist attractions, or points of interest (*POIs*), which include museums, parks, monuments, view points, castles, and the like. Our goal is to produce a short

*Research was conducted while working at Yahoo Research.

[1]Oxford dictionary defines a tip as "a small but useful piece of practical advice."

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052632

list of high-quality tips per POI, which can be consumed by the user in a short time and when using a small-screen device, such as a mobile phone. We therefore aim at high precision, i.e., high likelihood of a tip to be useful, even at the expense of recall, i.e, our ability to extract all tips unlocked in user reviews.

Our tip candidate extraction is based on n-gram templates. To increase generalizability, we allow the n-grams to include a one-word wildcard. The templates are derived from a gold set – the TripAdvisor city guide tips – based on recurring use. Applying these templates over user reviews, we obtain a set of tip candidates. To further increase precision, we rank the tip candidates for each POI and select the top 3. Ranking is performed by predicting a candidate's usefulness based on features of its text, the template(s) from which it was derived, the originating review, and the reviewer. In addition, tips that are very similar to those ranked higher are filtered out, to reduce redundancy.

For evaluation, we created two main datasets. The first is used for training and validation and the second for testing our entire pipeline. The tips in these datasets were annotated by an internal team of professional editors, who were educated and trained for the task of judging tip candidates. Our approach achieves high precision with decent coverage of POIs and cities, accounting for the more likely-to-be-viewed cities and POIs. The tip's text features demonstrated the most predictive power for usefulness, with the rest of the features adding little to nothing to performance.

Overall, our work offers the following key contributions:

- To the best of our knowledge, this is the first work to suggest tip extraction from user-generated reviews.
- To the best of our knowledge, we are the first to address the challenge of tip ranking.
- We present an extensive evaluation showing that our tips reach high precision, with good city and POI coverage.

2. RELATED WORK

Three approaches were proposed in the literature to improve user experience and decision making process in light of a constantly increasing number of available reviews [19]. The first is review ranking. The goal of this approach is creating a ranked list of reviews by independently assigning each review with a score expressing its quality and usefulness to the user [15, 23]. The second approach is review selection, which aims at identifying a subset of the most helpful non-redundant reviews that cover as many aspects as possible [20, 38, 19, 29]. The third approach is review summarization, in which review sentences expressing negative or positive opinions about different aspects are extracted and classified [14, 37, 8]. In contrast to the first two approaches, our approach operates at the review sentence level rather than the entire review. Different from the third approach, which also works at the sentence level, the tips we extract do not necessarily cover all possible aspects or express a positive or negative opinion. In a similar vain, multi-document summarization techniques can be applied to generate concise summaries of a review collection [9, 34, 10]. Yet, these summaries do not necessarily contain tips.

Somewhat related to tip extraction are studies about extraction of experiences [33, 26, 28, 36] and actions or todo items [35, 27, 32] using linguistic features. For example, Park et al. [33] defined "experience" as an activity or event an individual or a group has undergone, and classified experience-revealing sentences in blogs based on features such as tense, mood, aspect, and modality. Ryu et al. [35] detected "actionable clauses" in how-to instructions using linguistic features, including syntactic and modal characteristics. As for experiences, in this work we are interested in the more practical part of the content, which may leave out experience descriptions. On the other hand, the extracted tips do not necessarily include explicit instructions or commands of the type extracted for actions [35]; for example, a tip can also look like "Entrance is free on the last Friday of each month."

Another related body of research has focused on detecting advice-revealing text units. Wicaksono and Myaeng [41] proposed to use sequence models, such as conditional random fields, to extract advice sentences and respective context sentences from forum entries. Since the forum entries were not explicitly connected to a specific item, context such as when and where people may find the advice helpful was necessary. An earlier short paper by the same authors focused on the challenge of identifying advice-revealing sentences, working with data from travel blogs [40]. Various linguistic features defined by hand-crafted rules were used, including the appearance of terms such as "I suggest", "I strongly recommend", or "advice", with an associated proper noun, representing a travel entity, such as a hotel or a POI. Our approach, in contrast, automatically extracts templates from a seed of ground-truth data. Their evaluation was based on 207 blog posts whose sentences were manually labeled by the authors. The presence of an imperative mood expressions was found to be the most important feature. This imperative mood can be identified by the appearance of a verb in its simple unconjugated form, sometimes preceded by an adverb, at the beginning of the sentence. As part of our analysis, we examine the use of such imperative moods for candidate extraction. Kozawa et al. [18] studied the extraction of "prior" advice from the Web that can help plan outdoor activities. This study focused on the specific activities of climbing Mount Fuji and Mount Hotaka and was based on retrieved Web search results in Japanese. The features placed special emphasis on the end of the sentence, which carries high importance in Japanese.

Perhaps most closely related to ours is the work by Weber et al. [39], in which tips from Yahoo Answers were extracted to address Web search queries with "how-to" intent. We follow their definition of tips as "short, concrete and self-contained bits of non-obvious advice". Their tip extraction largely relied on the question-answer structure: only questions that start with "how to", "how do I", or "how can I" were considered. In addition, only best answers that were short enough and started with a verb were considered. The tip was always of the form "X:Y", where X is the tip's goal, taken from the question, and Y is the tip's suggestion, taken as the answer. For our purposes, this method is not fully applicable since neither questions nor queries are involved. Yet, as part of our experiments, we compare our template-based candidate extraction with the extraction of sentences that start with verbs.

3. TIP EXTRACTION

In this study, we use various datasets for extraction, training, and evaluation. Table 1 lists these datasets, which will

Table 1: Datasets used in this work.

	Size	Description
TAReviews	3,362,296 reviews	Publicly available TripAdvisor reviews for POIs in major cities within the U.S.
TAGuides	9,847 tips	City guide tips written by experts on TripAdvisor.
TipCands	413,125 sentences	Sentences longer than three words in TAReviews that match at least one template.
TipExtract	7,500 sentences	Sentences sampled from TipCands promoting balanced coverage of tip length, POI popularity, and templates.
TipRank	3,000 sentences	Top 3 ranked sentences from TipCands with respect to 1000 POIs.

be introduced in detail throughout the following sections. The first dataset is *TAReviews*, which includes publicly-available TripAdvisor reviews written in the years 2012-2015 for POIs in major U.S. cities. All cities with at least 50 POIs on TripAdvisor at the time of the study – 415 in total – were considered. Overall, TAReviews includes over 3.3M reviews of 20,335 POIs, with a total of 16.2M sentences. While examining the reviews, we asserted that a tip should consist of at least 4 words, as shorter sentences are not informative enough to serve as tips. TAReviews includes 15.4M sentences of 4 words or more.

3.1 Candidate Evaluation

Our evaluation of sentences as potential tips is based on manual labeling by in-house professional editors. The pool included a total of 30 editors, of whom different subsets were selected for different tasks, proportionally to the task's size, as detailed below. Unless otherwise stated, each sentence was evaluated by a single editor.

For each sentence, editors were presented with the sentence itself, as well as the city and POI it refers to. Our main evaluation criterion for a sentence was whether it represents a *useful* tip. Editors were told that a useful tip must always be clear, self-contained, and non-trivial, so that it can be presented to a user who plans a visit to the corresponding POI. In addition, we also set out to examine a few finer-grained aspects. The first was whether the sentence represents an *actionable* tip, i.e., a tip that allows the traveler to better prepare for the trip, e.g., to equip oneself with certain items or plan to arrive at certain days or hours. A sentence can represent both a useful and actionable tip (e.g., "The shows are not at regular hours, so make sure to check the schedule in advance"), useful but not actionable (e.g., "A fun place to take young kids and spend a rainy afternoon"), actionable but not useful (rarely, e.g., for an outdated tip, such as "Make sure to bring the coupons for the 2014 event"), and neither useful nor actionable (e.g., "This is a great place to visit").

The second aspect, for useful tips only, was whether they are useful *before* the trip (i.e., while planning the visit) or *during* the trip (i.e., while visiting). For example, a tip such as "Tours must be booked at least a week in advance" is useful before the trip, while a tip such as "Don't miss the Bamboo sculpture right before the Turtle Island" is useful during the trip. Some tips might be useful for both, e.g., "Free parking is only available at the harbor, with a 15 minute walking distance".

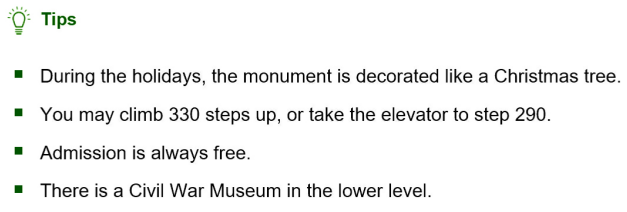

Figure 1: City guide tips on TripAdvisor.

Finally, to understand the reasons that make tips not useful, we composed a list of eight reasons we observed by inspecting TripAdvisor reviews: 'General/Trivial', 'Missing Context', 'Too Specific' (likely to be relevant only for a very specific time, event, or group of users), 'Outdated', 'Poor Language' (many spelling mistakes or a broken sentence), 'Spam' (heavily-suspicious advertisement), 'Offensive', and 'Other'. Editors were asked to select one of these reasons for tips marked as not useful.

To assess the difficulty of labeling sentences as useful and actionable, 10 editors labeled the same set of 100 random tip candidates. The inter-annotator agreement computed using Cohen's kappa [4] was 0.88 and 0.59 for useful and actionable tips, respectively. This indicates that the notion of usefulness is clear and reaches a high consensus, while the judgment of actionability is more controversial. Most of the following analyses focus on tip usefulness, assuming that useful tips are worth presenting even when they are not actionable.

3.2 Tip Extraction Methods

Gold Set. TripAdvisor features expert-authored travel guides for cities. Many of the guides include designated tips, manually entered by the city guide author for different POIs in the city, as demonstrated in Figure 1. These short tips, typically of one sentence, serve as our gold set for the type of tips we pursue by automatic extraction. We collected all city guide tips for POIs in U.S. cities that were available at the time of our study. Overall, we extracted 9,847 tips across 3,256 POIs within 54 different cities (all included in the 415 cities that are part of TAReviews), to create the *TAGuides* dataset. The city with the highest number of tips was New York City, with 689 tips across 225 POIs, while Houston had the smallest number of tips out of the 54 cities – 31, across 19 POIs. The POI with the highest number of tips was Central Park in New York City with 29 tips, while 1,291 POIs (39.6%) had only 1 tip. The TAGuides tips were evaluated by 30 editors; overall, 91.1% were marked useful. We only considered the city guide tips marked useful as part of our gold set.

Random Sentences. To get a sense of what portion of the sentences in TAReviews are good tip candidates, we examined 400 randomly sampled sentences of 4 words or more. For sentence splitting, we used the OpenNLP Sentence Detector[2]. The sentences were evaluated by 4 editors. Overall, 23.3% of the sentences were marked as useful and 11.5% as actionable. Sentences that were marked as not useful included details about the POI (e.g., "The store is in the

[2]`http://opennlp.apache.org`

front, the factory in the back – both are in what appears to be the original building from 1909"), opinions about the POI (e.g., "I have been to several art museums and have seen better ones"), descriptions of the overall experience (e.g., "We were pleasantly surprised at how much we all enjoyed the museum"), and personal details (e.g., "We got married in a Rose Garden so when we travel we always like to visit them").

Sentences Starting with a Verb. To improve the quality of tip candidates, we considered sentences starting with an infinitive verb [39]. A similar approach was also used to extract imperative moods – the most powerful feature found for detecting advice-revealing sentences within blog posts [40]. To identify infinitive verbs, we used OpenNLP part-of-speech (POS) tagger. We considered sentences starting with a verb in its base form (VB), a verb in its non-3rd person singular present form (VBP), an adverb followed by a base-form verb (RB+VB; e.g., "definitely take"), and an adverb followed by a non-3rd singular present verb (RB+VBP). We randomly sampled 400 sentences out of all sentences in TAReviews (with 4 words or more) meeting one of the above four constraints. These sentences were labeled as before by four annotators, with 48.3% marked useful (35.6% actionable). The vast majority (75.8%) of the sentences in this sample start with a base-form verb (VB). Considering only these sentences, the portion of useful tips rises to 53.5% (39.4% actionable). A similar precision (52.3%) was reported for the extraction of advice-revealing sentences and was declared as "far short of the ideal situation" [40]. Finally, it should be noted that a high portion of the TAGuides tips do not start with a verb: only 23.5% start with a VB.

Templates. To develop a more precise candidate extraction method, we used our gold set of TAGuides tips marked useful by editors. We set out to understand what language characterizes high-quality tips of the type entered manually by subject matter experts. To this end, we sought for n-grams that repeat in a large number of tips. Specifically, we considered n-grams with $4 \leqslant n \leqslant 7$. Larger values of n produced term sequences that rarely recurred, while smaller values of n produced too generic sequences. To enable some level of generalization, we allowed one word of a given n-gram to be a *wildcard*, which may represent any single word. The wildcard could be positioned anywhere within the n-gram (first, last, or in the middle). For example, in the n-gram "be sure to $*$ the", the wildcard, marked by '$*$', can stand for 'visit', 'check', 'see', 'ask', 'watch', and so forth; in "check out the $*$", the wildcard may stand for 'website', 'gift', 'special', etc.

For each $4 \leqslant n \leqslant 7$, we extracted a list of 1-wildcard n-grams, henceforth *templates*, from the gold set of the city guide tips. We manually filtered out templates that were too generic or irrelevant, such as "the $*$ of the". We also set a minimum threshold over the number of tips each template occurs in, e.g., for $n=7$ we required that the template occurs in at least 3 tips. Finally, we required that the template's wildcard captures at least two different words across its occurrences. Overall, we produced a list of 150 templates, whose statistics is detailed in Table 2. The template with the highest number of occurrences is "be sure to $*$", occurring in 359 tips, with 83 different wildcard words[3].

Table 2: Template statistics: (i) the number of templates of size n ('Templates'); (ii) the total ('Total'), median ('Med'), minimum ('Min'), and maximum ('Max') number of tips matching the templates of size n; and (iii) the median, minimum, and maximum number of unique wildcard words per value of n.

		Tips				Wildacrds		
n	Templates	Total	Med	Min	Max	Med	Min	Max
4	21	1,645	70	32	359	16.5	3	131
5	45	1,220	21	12	79	8	2	71
6	50	517	7.5	6	49	4	2	33
7	34	214	5	4	39	3	2	32

4. TIP CHARACTERISTICS

We applied the set of 150 templates on the set of reviews in TAReviews to create a dataset of tip candidates – *TipCands*. Overall, TipCands includes tip candidates for 396 cities out of the 415 (95.4%), compared to only 54 cities (13%) included in TAGuides. The tips cover 15,198 POIs (62% of all relevant POIs), compared to 3,256 (13%) POIs covered by TAGuides. On average, the number of tips per city is 1,045 (median: 364). New York City has the highest number of tips (27,979 across 369 POIs), followed by Orlando (15,926 tips, 72 POIs) and San Francisco (14,018 tips, 201 POIs). The city with the lowest number of tips is Yakima, WA, with 23 tips. The average number of tips per POI is 27.1 (median: 5), with 35.5% having 10 tips or more. The POI with the highest number of tips is Central Park in New York City (4,117 tips), as in TAGuides.

The cities and POIs not covered by the extracted tips naturally tend to the uncommon; we therefore expect this coverage to allow handling the vast majority of user demand in a travel application. We now turn to examine the quality and characteristics of the extracted tip candidates.

4.1 Useful vs. Not Useful Tip Candidates

We sampled 7,500 tips from TipCands, ensuring fair representation of tip length, POI popularity, and matched templates. Specifically, we partitioned all the extracted tips into nine buckets by (i) three equal thirds according to the tip length (by number of words), and (ii) three equal thirds according to the related POI's popularity (by number of reviews it received). Finally, we sampled from each bucket a set of tips that spans the different templates as equally as possible. This set of tips, henceforth the *TipExtract* dataset, was evaluated by 25 editors.

Table 3 presents the evaluation results for TipExtract. Overall, 73.2% of the tips were marked useful and 51.4% actionable; about two thirds of the useful tips were marked actionable. The majority of the tips were found useful before and not during the trip. Table 3 also shows the distribution of selected reasons for not-useful tips. The majority were general or trivial (e.g., "A nice way to spend the day"), while only a minority used poor language, were offensive, or suspicious of spamming. While the presentation of general

[3]The full list of templates is available at `http://www.ise.bgu.ac.il/downloadMe/templates.txt`

Table 3: Editorial results for TipExtract and TipRank.

		TipExtract	TipRank
Useful & Actionable	Useful	73.2	90.1
	Actionable (out of all tips)	51.4	75.0
	Actionable (out of all useful tips)	67.2	81.1
Before & During	Useful before the trip	68.3	61.6
	Useful during the trip	20.4	20.4
	Useful before and during the trip	11.3	18.0
Not Useful Reasons	General/Trivial	61.1	73.0
	Missing Context	13.7	8.4
	Too Specific	8.4	7.7
	Outdated	5.0	6.9
	Poor Language	2.7	1.8
	Spam	0.8	0.7
	Offensive	0.7	0.4
	Other	7.6	1.1

Table 4: Useful tip examples.

Harvard University	"If you have to drive there, make sure to book online for one of their parking garages before you get there."
Museum of Indian Arts & Culture, Santa Fe	"Do not miss the small enclosed sculpture garden as you approach the building!"
Houston Zoo	"The first Tuesday of each month (Sept.-May) is free."
Kapoho Tide Pools, Island of Hawaii	"It's a good idea to snorkel with water shoes on instead of fins just in case you need to climb over the lava to get from pool to pool."
Key West Lighthouse	"Be sure to go early in the day as it gets pretty hot climbing the 88 stairs to the top of the lighthouse."

or trivial tips is not likely to add any value, they are not disturbing or harmful such as spam, tips with poor language use, or even missing context. Table 4 lists a few examples of TipCands tips that were marked useful.

To further study the differences between useful and not-useful tip candidates, we examined their characteristics spanning 4 categories: the tip itself, the template(s) by which it was extracted, the originating review, and the authoring reviewer. Table 5 shows the statistics for various attributes across these categories.

Tip. Inspecting tips by their length in words, it appears that shorter tips, of up to 15 words, are less likely to be useful, probably as they more often tend to be general or trivial. Beyond 15 words, the differences are not large, but the "optimal" length seems to be $25-29$ words. We also examined the PCFG parse score of tips by applying syntactic parsing using the Stanford parser [17] and bucketing the tip candidates by their length-normalized log probability score. Sentences with higher (less negative) parse score are more likely to be not-useful tips. One explanation may be that sentences that contain relatively rare words and entity names result in a lower score, while those that use only common words (such as 'great' or 'city') get a higher score, yet often lack a meaningful or unique advice.

Template. The portion of useful tips produced by templates of length 4 to 6 was very similar, while substantially higher for templates of length 7.[4] As aforementioned, we considered all templates of length 7 that were matched by at least 3 city guide tips. Considering templates of higher length and/or with lower number of matching tips may lead to the extraction of over-specialized tips. Some increase can also be observed in the portion of useful tips as the number of matched templates grows.

Review. It can be seen that tips extracted from longer reviews are somewhat more likely to be useful. We also noticed that tip candidates that appear as the first sentence of a multi-sentence review are less likely to be useful (17% of the candidates account for these, with less than 60% marked useful). It could be that the opening sentence tends to convey a general description or a personal experience. Low review rating (3 stars and especially 2 or 1) is also a strong signal for lower usefulness likelihood; yet, the vast majority of reviews have ratings of 4 or 5, for which the useful po-

Table 5: Tip characteristics. 'R', '%' and '%U' denote the range of value buckets, the distribution of tips across buckets, and the portion of useful tips in a bucket, respectively.

Tip	Length (Number of Words)							
	R	$5-9$	$10-14$	$15-19$	$20-24$	$25-29$	$30-34$	35+
	%	11.7	27.2	24.6	16.5	9.6	5.1	5.4
	%U	60.2	67.4	76.1	78.9	82.8	76.2	80.6
	PCFG Parse Score							
	R	<(-7.4)	(-7.4) – (-6.9)		(-6.9) – (-6.5)		(-6.5) – (-6.1)	>(-6.1)
	%	20.0	20.0		20.0		20.0	20.0
	%U	81.0	80.0		77.4		69.0	59.5
Template	Length (n)							
	R		4		5		6	7
	%		20.3		36.5		26.1	17.1
	%U		72.0		71.8		71.7	80.3
	Number of Matched Templates							
	R		1		$2-3$		$4-5$	6+
	%		37.7		30.0		13.1	19.3
	%U		71.2		72.9		78.6	73.9
Review	Length (Number of Sentences)							
	R		1		$2-3$		$4-5$	$6-9$	10+
	%		2.3		22.0		25.3	28.9	21.5
	%U		73.5		69.3		71.3	74.7	77.5
	Rating							
	R		1		2		3	4	5
	%		0.6		1.0		5.8	27.9	64.7
	%U		31.9		49.3		66.4	74.0	74.3
Reviewer	TripAdvisor Level								
	R	0	1	2	3	4	5	6
	%	3.0	2.4	6.3	15.9	18.6	19.5	34.3
	%U	67.5	69.0	72.7	73.6	71.9	74.0	74.0
	Number of Helpful Votes							
	R	0	$1-2$	$3-5$	$6-10$	$11-50$	$51-100$	101+
	%	2.8	8.5	8.4	11.1	40.5	13.7	14.9
	%U	73.2	73.1	70.1	72.4	73.1	74.2	74.7

[4]For this analysis, we considered the length of the longest template the tip matched.

tions are similar. The tendency of user-review ratings to the positive has indeed been often observed [6].

Reviewer. Our analysis indicates that the majority of reviews originate from experienced reviewers, with considerable numbers of authored reviews, granted helpful votes, and high TripAdvisor level[5]. A low TripAdvisor level (0 or 1) is indicative of lower usefulness likelihood, but covers a small portion of the reviewers. A similar trend was observed when inspecting the reviewer's number of total reviews. For helpful votes, however, we could not observe any notable trend[6].

In addition to the characteristics presented in Table 5, we inspected other aspects related to the templates and the tip's text. Table 6 shows, for each $4 \leq n \leq 7$, the top and bottom template by the portion of useful tips it produced. While the top templates capture concrete aspects such as arrival time, fee charging, and opening hours, the bottom ones are more general in nature. Indeed, there was a negative correlation between the number of tips captured by a template and the template's portion of useful tips ($r = -0.4, p < 0.01$). While the template "this is a great place to $*$" may often capture a general recommendation about the POI, which is not a tip, in other cases it captures more specific aspects, e.g., "This is a great place to see a wide variety of birds, especially in the spring and fall when they're migrating". Finally, it is worth noting the high correlation between the number of tips captured by a template and the number of occurrences of the template in the city guide tips ($r = 0.75, p < 0.01$).

We also set out to more closely examine the language of the tip. To this end, we considered lexical features, i.e., the appearance of specific terms. Table 7 shows a few examples of the most distinctive unigrams for useful and not-useful tip candidates. For distinctive terms, we calculated those that contribute the most to the KL divergence between the language model induced from useful tips and that induced from not-useful tips and vice versa [2, 12][7]. It can be seen that terms such as 'early', 'free', or 'reservation' are especially characteristic of useful tips, while terms such as 'great' and 'place', which are more general in essence, or 'was', which is often used to describe a past experience or historical fact, are more typical for not-useful tips.

A similar analysis for bigrams and trigrams reveals the more fine-grained structure of distinctive terms. For example, the verb 'make' is among the distinctive useful unigrams (beyond the ones presented in Table 7), with "make sure" among the top useful bigrams and "make time" among the top not-useful bigrams. Similarly, "check the website" is on the list of useful trigrams, while "check it out" is on the not-useful list. As another example, each of the unigrams in the trigram "bring your camera" is on the top useful list, but the trigram itself is on the not-useful list.

To generalize from semantics to syntax, we applied part-of-speech tagging over the useful and not-useful tips. Table 7 shows the most distinctive POS tags, based on KL divergence over the resultant POS tags language models. For example, past-tense verbs (VBD) and personal pronouns (PRP) are among the highest on the not-useful list, as they are often used to describe personal experiences. On the other hand, adverbs (RB) and non-3rd person singular present verbs (VBP) are among the highest on the useful list.

[5]Level ranges from 0 to 6 based on a contribution-driven point system.

[6]We also examined the average number of helpful votes per review for each reviewer, but no signal was found.

[7]Add-one smoothing was used.

Table 6: Top and bottom templates of length n by portion of useful tips ('%U'). In addition, the number of tips that match the template ('Tips'), the number of unique wildcard words ('Wildcards'), and the portion of actionable tips ('%A').

	n	Template	Tips	Wildcards	%U	%A
Top	4	make a reservation $*$	1,091	117	94	94
	5	go early in the $*$	2,237	36	97	97
	6	there is a $*$ fee to	522	56	95	77
	7	the $*$ is open 7days a week	25	15	100	94
Bottom	4	you want to $*$	49,538	1,279	63	45
	5	is one of the $*$	21,631	1,031	46	15
	6	a great way to $*$ the	4,657	188	48	10
	7	this is a great place to $*$	8,749	327	61	10

Table 7: Most distinctive unigram terms and POS tags characterizing useful and not-useful tips. POS tags are presented using the standard Penn Treebank notation [24].

Unigrams		POS tags	
Useful	Not Useful	Useful	Not Useful
early	great	NNS	DT
free	place	IN	TO
avoid	want	EX	VBZ
reservation	way	RB	VBD
tours	was	VBP	PRP

5. TIP CLASSIFICATION

After gaining a better understanding of the characteristics of useful versus not-useful tip candidates, we set out to build a classifier to distinguish between them.

Setup. We designed the classification features following the analysis presented in Section 4, spanning the same four categories: (i) the tip's text, (ii) the matched templates, (iii) the originating review, and (iv) the reviewer. The features derived from the tip's text can be further divided into four sub-categories: (i) lexical features, (ii) word embedding, (iii) POS tags, and (iv) text surface and quality descriptors. Table 8 presents a detailed description of the features.

We experimented with three classifiers: Logistic Regression, AROW (Adaptive Regularization of Weights [7]), and SVM (Support Vector Machine)[8]. We used 10-fold cross-validation to tune the hyper-parameters and evaluate the classifiers. As the evaluation metric, we used the area under the ROC curve (AUC) rather than accuracy, as our final goal is to produce a ranked list of tips.

For Logistic Regression, we used stochastic gradient descent with the ADAM method for adaptive learning rate [16] and tuned the initial learning rate, the regularization weight, and the number of training iterations. For AROW, we tuned the r hyper-parameter and the number of training iterations. We used SVM with a linear kernel and tuned the error penalty hyper-parameter C. In addition, as our dataset is imbalanced, containing only 26.8% negative examples,

[8]For Logistic Regression and AROW, we used an internal implementation; for SVM we used LIBSVM [5].

Table 8: Features used for tip classification and ranking.

Tip	TGrams	Number of *term* occurrences in the tip: unigrams (UTGrams), bigrams (BTGrams), and trigrams (TTGrams). We consider uni/bi/tri-grams appearing at least m times in the training set. We set m=5 to optimize average AUC via 10-fold cross-validation over the training data following experiments with $m \in \{1, 3, 5, 7, 10\}$.
	W2V	Weighted centroid representation of the tip using Word2Vec [25]. Each term in the tip is represented as a vector of 300 dimensions. The vectors are weighted uniformly (UniW2V) or by the inverse document frequency (IDFW2V) of the corresponding term. The TAReviews dataset was used to train Word2Vec vectors and compute IDF values.
	POSGrams	Number of occurrences of *part-of-speech* unigrams, bigrams, and trigrams in the tip. Uni/bi/tri-grams appearing less than m=5 times in the training set were discarded, as was the case with TGrams.
	Surface	– Tip length in words. – Punctuation marks and capital letters: counts and normalized ratios to tip length in characters. – Length-normalized log probability of the PCFG parse score.
Template		– IDs of all matched templates. – Number of templates that the tip matches. – Wildcard positions (first, middle, or last), aggregated across all matched templates. – Number of terms (n) in the longest template the tip matches.
Review		– Review length: number of sentences, words, and ratio of words to sentences. – Tip position within the review (first, middle, or last) for multi-sentence reviews. – Rating (from 1 to 5) assigned by the reviewer. – Number of "helpful" votes the review received from the community.
Reviewer		– Number of reviews written by the reviewer. – Number of "helpful" votes across all reviews authored by the reviewer. – The ratio between the number of "helpful" votes and the total number of reviews. – TripAdvisor level (from 0 to 6).

Table 9: AUC performance of Logistic Regression, AROW, and SVM using the full feature set and the best performing feature subset.

	All Features	Best Feature Subset
Logistic Regression	0.792	0.792 (All features)
AROW	0.770	0.778 (Excluding POSGrams)
SVM	0.713	0.770 (Using only W2V)

we evaluated whether over-sampling the negative examples would improve the results[9]. We considered negative-to-positive example ratios ranging from 1:1 to 4:1. For Logistic Regression, we found that a 2:1 ratio yields the best AUC, while for AROW it was a 3:1 ratio. For SVM, the error penalty C can be set differently for each class to achieve the same effect. We experimented with negative-to-positive penalty ratios from 1:1 to 4:1 and found that a 2:1 ratio yields the best AUC.

Results. Table 9 presents the AUC performance of the three classifiers when using all the features and the best subset of features. The best performance is attained by Logistic Regression when using all features, followed by AROW when excluding POSGrams. To better understand the contribution of the different features, we trained the Logistic Regression classifier with different feature subsets. Table 10 presents the results.

Using the lexical unigrams (UTGrams) alone yields fairly high performance; yet, adding bigrams (BTGrams) and trigrams (TTGrams) further improves the performance; as seen in Section 4.1, bigrams and trigrams capture finer-grained meaning and further distinguish the language of useful from not-useful tips. Inspecting the two types of Word2Vec-weighted centroids, we see that the uniform-weighted centroid (UniW2V) yields a substantially higher performance than the IDF-weighted centroid (IDFW2V). This may suggest that for our task, the frequent (low IDF) terms, are as important as the less frequent (high IDF) terms. In particular, many of the frequent terms may stem from the templates themselves. Overall, the combination of both centroids (W2V) achieves similar and even slightly higher performance than the lexical features (TGrams). Combining both of these feature subsets (TGrams + W2V) further improves the performance. The POSGrams features, which have been shown most effective in previous studies [39, 40, 41], achieve much lower performance than W2V and TGrams. Moreover, adding them on top of TGrams + W2V does not improve the performance. Apparently, TGrams and W2V already capture the relevant discriminative information conveyed in POSGrams, but add finer-grained distinctions thus leading to higher performance. It is worth mentioning that the most important POSGrams feature (across unigrams, bigrams, and trigrams) was "ST-VB", which indicates a sentence that starts with a base-form verb. This coincides with previous work [39, 40] and with our findings from Section 3. The general text features (Surface) capture some useful information on their own; when added to the TGrams and W2V features, the performance is slightly improved.

Inspecting feature subsets beyond the tip itself, it can be seen that the template features attain fairly good performance on their own, however their exclusion does not substantially degrade performance. The review features achieve lower results on their own, and their exclusion decreases performance only by a small extent. Finally, the reviewer features do not pose any contribution. Overall, these results indicate that for classifying useful and not-useful tip candidates, the tip features are by far the most informative.

It should be mentioned that the same analysis with the AROW classifier yielded similar results, with the TGrams and W2V features achieving the highest performance, and their combination further improving the results. As shown in Table 9, for AROW the exclusion of POSGrams improved the performance to some extent.

[9]Over-sampling was performed only on the training folds, while the test folds were evaluated using their original imbalanced distribution.

Table 10: AUC performance when using ('Only') or removing ('Exclude') subsets of features when training the Logistic Regression classifier.

Features	Only	Exclude
UTGrams	0.737	0.791
UTGrams + BTGrams	0.758	0.788
TGrams = UTGrams + BTGrams + TTGrams	0.763	0.785
UniW2V	0.762	0.787
IDFW2V	0.730	0.787
W2V = UniW2V + IDFW2V	0.772	0.774
POSGrams	0.713	0.792
Surface	0.601	0.790
TGrams + W2V	0.789	0.743
TGrams + POSGrams	0.767	0.785
TGrams + Surface	0.765	0.785
TGrams + W2V + POSGrams	0.788	0.743
TGrams + W2V + Surface	0.791	0.740
Tip = TGrams + W2V + POSGrams + Surface	0.790	0.740
Template	0.689	0.791
Review	0.587	0.790
Reviewer	0.496	0.791
All	0.792	–

6. TIP RANKING

To further improve the quality of extracted tips, we set out to rank the tips for POIs with more than 3 tips. To this end, we used the classifier's output score to rank all the POI's tip candidates by their likelihood to be useful. This approach allowed us to optimize the selection of the top-3 tips per POI and reflects our goal of providing a small list of high-quality tips that the user can quickly consume.

Setup. To evaluate the ranking approach, we created another dataset, denoted *TipRank*, in which lists of top-3 tips were included for different POIs. For this dataset, we only considered POIs that had at lest 4 tips in TipCands (56.9% of the POIs), since for the rest, ranking cannot improve the quality of their top-3 tips. The set of tips from these POIs still covers all cities and 96.7% of all tips in TipCands. Specifically, we sampled 1,000 such POIs from TipCands, weighted by their popularity (number of reviews in TAReviews), so as to reflect the likelihood of a POI's tips to be presented to a user[10]. For each selected POI, we ranked all its extracted tips according to the classifier's score. The entire TipExtract dataset was used to train the classifier using the optimal hyper-parameter values found by the 10-fold cross-validation, as described in Section 5. For various production considerations, we used AROW (with all features but POSGrams) as our classifier.

Filtering Similar Tips. Inspecting the resultant tip rankings per POI, we noticed that some tips provide very similar information. For example, the two tips "It's a great place to go snorkeling, especially during low tide and calm days" and "Be sure to go at low tide, as you really can't snorkel at high tide" convey the same message, hence showing both is not desirable. We therefore set out to remove very similar, or *redundant*, tips. To this end, we scanned the ranked list of tips from top to bottom and filtered out tips whose similarity to one of the preceding tips (if there were any) was higher than a predefined threshold θ, until 3 tips were accumulated.

To compute the similarity between tips t_i and t_j, we used a linear interpolation [22] of the word order (WO) and semantic (SEM [31]) similarity measures based on Word2Vec word representations:

$$SIM(t_i,t_j)=(1-\alpha)WO(t_i,t_j)+\alpha\sqrt{SEM(t_i,t_j)SEM(t_j,t_i)},$$

where α is a free parameter, set to 0.85 in our experiments, as in [22]. This interpolation was found to be the most effective for measuring sentence similarity compared to a variety of alternative measures [1]. The SEM measure was adapted for Word2Vec representation as follows:

$$SEM(t_i,t_j)=\frac{\sum_{w_i\in t_i}\max_{w_j\in t_j}COS(W2V(w_i),W2V(w_j))IDF(w_i)}{\sum_{w_i\in t_i}IDF(w_i)};$$

$COS(W2V(w_i),W2V(w_j))$ is the cosine between the vectors representing words w_i and w_j, which was also used for estimating word similarity in the WO measure [22]; $IDF(w)$ is the inverse document frequency of term w.

We experimented with 3 tip filtering levels by splitting (at random) the set of 1,000 POIs in TipRank into 3 equal-size sets. For each set, a different value of θ was used. Specifically, we selected values for θ so as to filter out 0% (i.e., no filtering), 10%, and 20% of the tips in the top 3, corresponding to θ=1.0, θ=0.75, and θ=0.69, respectively.

The three sets of tips were evaluated by 12 editors. All 3 tips for each POI were evaluated successively by the same editor. In addition to the previous instructions, editors were asked to judge whether a tip is redundant given the preceding tips presented for the same POI. A tip is considered redundant if the information it provides was already covered by the tips ranked above. We further distinguished between two cases: (i) the tip is contained in or equivalent to a preceding tip, and (ii) only part of the tip is covered by its preceding tips. In the latter case, editors were asked to judge whether the extra information is useful on its own.

Results. Table 3 presents the results for the TipRank dataset alongside those of the TipExtract dataset. The overall portion of useful tips is 90.1%, an increase of over 23% compared to the TipExtract dataset and comparable to the portion of useful tips in TAGuides (91.1%, as reported in Section 3).

Interestingly, the portion of actionable tips also substantially increased, although this was not explicitly included in the objective of our training process. Out of the useful tips, the portion of tips useful both before and during the trip increased at the expense of tips useful only before the trip, while the portion of tips useful only during the trip remained the same. Inspecting the distribution of reasons for labeling tips as not useful, a substantial increase in the portion of the 'General/Trivial' tips can be observed. The sharpest decrease was in the 'Other' and 'Missing Context' tips. The latter can be a result of the classifier penalizing sentences starting with terms such as 'so', 'but', and 'then', which imply that the tip is connected to the previous sentence. Finally, reflecting on our analysis in Section 3, it is worth mentioning that only 9.8% of the 3,000 tips in the TipRank dataset start with a base-form verb.

Table 11 presents the portion of useful and redundant tips for the three POI sets according to the similarity threshold θ. Overall, the portion of useful tips in each set is rather

[10]POIs included in TipExtract were discarded.

Table 11: Portion of useful and redundant tips when using different thresholds for tip filtering. 'All' and 'Shortest': tips of all lengths and the 25% shortest tips, respectively.

	All		Shortest	
Filtered Tips	Useful	Redundant	Useful	Redundant
0% (θ=1.00)	91.9	3.8	92.4	9.4
10% (θ=0.75)	88.3	3.5	87.8	5.4
20% (θ=0.69)	90.2	2.7	91.9	3.7

similar, with the set for which no tips were filtered out receiving the highest portion. This set also has, as expected, the highest portion of redundant tips. When applying our tip filtering procedure, the portions of redundant tips decrease to some extent.

The overall portion of redundant tips is not very high, but may become a more acute issue in two cases: (i) when selecting more than 3 tips, as each tip has higher likelihood of being similar to one of the preceding tips; and, (ii) when presenting shorter tips,as they convey less information. To gain further understanding of the second case, we inspected the 25% shortest tips in TipRank. As shown in Table 11, the portion of useful tips does not substantially change for short tips, but the portion of redundant tips is substantially higher if no filtering is performed. Applying our filtering procedure reduces the portion of redundant tips, without a substantial drop in the portion of useful tips.

7. DISCUSSION AND FUTURE WORK

We introduced tip extraction and ranking from user generated reviews. Our approach can complement existing methods for handling large review volumes, especially in scenarios where the user is preparing for a trip to a given destination and is looking for a few practical pieces of advice, beyond the "standard" descriptions. Our template-based tip extraction method demonstrated high precision, with over 73% of the tips marked as useful. Not only did this provide a good starting point for tip quality, but also enabled to focus the editors' efforts on a less noisy dataset. This method, however, also has drawbacks, such as (i) restricting tips to predefined patterns, and (ii) relying on a gold set of domain-specific tips. We tried to address the first issue by using different types and lengths of templates. Generalizing to other websites and domains is also worth exploring. For example, some of the templates may reflect common tip language across many domains, while others may be specific to travel or even just to POIs.

Our goal in this work was to provide a small set of high-quality tips per POI. We hence focused on precision rather than recall. The set of 150 templates yielded tips for 62% of the POIs and 75% of those with at least one review. Moreover, POIs with no extracted tips are among the less popular, thus also less likely to attract user interest. Overall, our extraction method produced a tip candidate per 8 reviews, while our initial evaluation of random review sentences showed a much higher potential. In cases where coverage becomes more important (e.g., when aiming to select higher number of tips per item), the set of templates may be further generalized to capture more tip candidates.

Ranking the top 3 tips per POI led to a substantial quality improvement for POIs with more than 3 extracted tips (which are also the more popular POIs), reaching a similar precision to the manually-entered city-guide tips. While the training process focused on increasing the portion of useful tips, the results also showed improvement in other aspects, including the portion of actionable tips, the portion of tips useful both before and during the trip, and the portion of tips marked not useful for merely being general or trivial. For other not-useful reasons, specialized methods can be applied to further reduce their occurrence, e.g., anaphora resolution [21] to detect missing context; enhanced date analysis to detect outdated tips; profanity filtering to avoid offensive language; or advanced methods for spam detection [13].

The key predictive features for tip usefulness were found to be based on the tip's text, with very minor contribution from review and reviewer features. Our analysis of tip characteristics indicated that these are often highly skewed (e.g., review rating, TripAdvisor level), and might therefore not be effective as discriminative features. As for the tip itself, POS tags, which were used as a key means for extracting tips or advice in previous studies [39, 40, 41], were not found to add any value on top of the lexical and Word2Vec features. This implies that the distinguishing characteristics of a useful tip lie in its semantics more than in its syntax.

Our tip similarity measure was helpful in filtering out redundant tips, especially short ones. Tip similarity can further be used for clustering similar tips per POI, which can help enhance tip presentation (e.g., by supporting a "more like this" feature) and ranking (e.g., by considering the cluster size). These ideas are left for future research.

8. REFERENCES

[1] P. Achananuparp, X. Hu, and X. Shen. The evaluation of sentence similarity measures. In *Proc. of DaWaK*, pages 305–316, 2008.

[2] A. Berger and J. Lafferty. Information retrieval as statistical translation. In *Proc. of SIGIR*, pages 222–229, 1999.

[3] S. Burgess and C. Sellitto. User-generated content (ugc) in tourism: Benefits and concerns of online consumers. In *Proc. of ECIS*, pages 417–429, 2009.

[4] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Comput. Linguist.*, 22(2):249–254, June 1996.

[5] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):1–27, 2011.

[6] J. A. Chevalier and D. Mayzlin. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research*, 43(3):345–354, 2006.

[7] K. Crammer, A. Kulesza, and M. Dredze. Adaptive regularization of weight vectors. *Machine Learning Journal*, 91(2):155–187, 2013.

[8] P. Cremonesi, R. Facendola, F. Garzotto, M. Guarnerio, M. Natali, and R. Pagano. Polarized review summarization as decision making tool. In *Proc. of AVI*, pages 355–356, 2014.

[9] G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *JAIR*, 22:457–479, 2004.

[10] M. Gambhir and V. Gupta. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, pages 1–66, 2016.

[11] U. Gretzel, K. H. Yoo, and M. Purifoy. Online travel review study: Role and impact of online travel reviews. 2007.

[12] I. Guy. Searching by talking: Analysis of voice queries on mobile web search. In *Proc. of SIGIR*, pages 35–44, 2016.

[13] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari. Detection of review spam: A survey. *Expert Systems with Applications*, 42(7):3634 – 3642, 2015.

[14] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proc. of KDD*, pages 168–177, 2004.

[15] S. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proc. of EMNLP*, pages 423–430, 2006.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[17] D. Klein and C. D. Manning. Accurate unlexicalized parsing. In *Proc. of ACL*, pages 423–430, 2003.

[18] S. Kozawa, M. Okamoto, S. Nagano, K. Cho, and S. Matsubara. Advice extraction from web for providing prior information concerning outdoor activities. In *Intelligent Interactive Multimedia Systems and Services*, pages 251–260. Springer, 2011.

[19] T. Lappas, M. Crovella, and E. Terzi. Selecting a characteristic set of reviews. In *Proc. of KDD*, pages 832–840, 2012.

[20] T. Lappas and D. Gunopulos. Efficient confident search in large review corpora. In *Proc. of ECML PKDD*, pages 195–210, 2010.

[21] S. Lappin and H. J. Leass. An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, 20(4):535–561, Dec. 1994.

[22] Y. Li, Z. Bandar, D. McLean, and J. O'Shea. A method for measuring sentence similarity and its application to conversational agents. In *Proc. of FLAIRS*, pages 820–825, 2004.

[23] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proc. of WWW*, pages 691–700, 2010.

[24] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, 1993.

[25] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. 2013.

[26] H.-J. Min and J. C. Park. Identifying helpful reviews based on customer's mentions about experiences. *Expert Systems with Applications*, 39(15):11830–11838, 2012.

[27] K. Nagao, K. Inoue, N. Morita, and S. Matsubara. Automatic extraction of task statements from structured meeting content. In *Proc. of IC3K*, volume 1, pages 307–315. IEEE, 2015.

[28] Q. Nguyen. *Detecting experience revealing sentences in product reviews*. PhD thesis, University of Amsterdam, 2012.

[29] T. Nguyen, H. W. Lauw, and P. Tsaparas. Using micro-reviews to select an efficient set of reviews. In *Proc. of CIKM*, pages 1067–1076, 2013.

[30] P. O'Connor. User-generated content and travel: A case study on tripadvisor.com. In *Proc. of ENTER*, pages 47–58, 2008.

[31] A. Omari, D. Carmel, O. Rokhlenko, and I. Szpektor. Novelty based ranking of human answers for community questions. In *Proc. of SIGIR*, pages 215–224, 2016.

[32] P. Pareti, E. Klein, and A. Barker. A semantic web of know-how: Linked data for community-centric tasks. In *Proc. of WWW Companion*, pages 1011–1016, 2014.

[33] K. C. Park, Y. Jeong, and S. H. Myaeng. Detecting experiences from weblogs. In *Proc. of ACL*, pages 1464–1472, 2010.

[34] D. R. Radev, H. Jing, M. Sty, and D. Tam. Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938, 2004.

[35] J. Ryu, Y. Jung, and S.-H. Myaeng. Actionable clause detection from non-imperative sentences in howto instructions: A step for actionable information extraction. In *Proc. of TSD*, pages 272–281. Springer, 2012.

[36] C. S. Sauer and T. Roth-Berghofer. Solution mining for specific contextualised problems: Towards an approach for experience mining. In *Proc. of WWW Companion*, pages 729–738, 2012.

[37] R. Sipos and T. Joachims. Generating comparative summaries from reviews. In *Proc. of CIKM*, pages 1853–1856, 2013.

[38] P. Tsaparas, A. Ntoulas, and E. Terzi. Selecting a comprehensive set of reviews. In *Proc. of KDD*, pages 168–176, 2011.

[39] I. Weber, A. Ukkonen, and A. Gionis. Answers, not links: Extracting tips from yahoo! answers to address how-to web queries. In *Proc. of WSDM*, pages 613–622, 2012.

[40] A. F. Wicaksono and S.-H. Myaeng. Mining advices from weblogs. In *Proc. of CIKM*, pages 2347–2350, 2012.

[41] A. F. Wicaksono and S.-H. Myaeng. Toward advice mining: Conditional random fields for extracting advice-revealing text units. In *Proc. of CIKM*, pages 2039–2048, 2013.

[42] Z. Xiang and U. Gretzel. Role of social media in online travel information search. *Tourism management*, 31(2):179–188, 2010.

[43] K. Yatani, M. Novati, A. Trusty, and K. N. Truong. Review spotlight: A user interface for summarizing user-generated reviews using adjective-noun word pairs. In *Proc. of CHI*, pages 1541–1550, 2011.

[44] Q. Ye, R. Law, and B. Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.

[45] Q. Ye, R. Law, B. Gu, and W. Chen. The influence of user-generated content on traveler behavior: An empirical investigation on the effects of e-word-of-mouth to hotel online bookings. *Comput. Hum. Behav.*, 27(2):634–639, Mar. 2011.