GOOD GRIEF, I CAN SPEAK IT!
PRELIMINARY EXPERIMENTS IN AUDIO RESTAURANT REVIEWS

Joseph Polifroni[1], *Stephanie Seneff*[2], *S.R.K. Branavan*[2], *Chao Wang*[3], *Regina Barzilay*[2]

[1]Nokia Research Center
4 Cambridge Center
Cambridge, MA 02142 USA

[2]MIT CSAIL
32 Vassar Street
Cambridge, MA 02139 USA

[3]Vlingo
17 Dunster Street
Cambridge, MA 02138 USA

joseph.polifroni@nokia.com, [seneff,branavan,regina]@csail.mit.edu, wangc@vlingo.com

ABSTRACT

In this paper, we introduce a new envisioned application for speech which allows users to enter restaurant reviews orally via their mobile device, and, at a later time, update a shared and growing database of consumer-provided information about restaurants. During the intervening period, a speech recognition and NLP based system has analyzed their audio recording both to extract key descriptive phrases and to compute sentiment ratings based on the evidence provided in the audio clip. We report here on our preliminary work moving towards this goal. Our experiments demonstrate that multi-aspect sentiment ranking works surprisingly well on speech output, even in the presence of recognition errors. We also present initial experiments on integrated sentence boundary detection and key phrase extraction from recognition output.

Index Terms— Speech applications, content creation, sentiment detection, speech summarization, user modelling

1. INTRODUCTION

A couple visiting Toronto have just finished a meal at a Chinese restaurant. As they walk out of the restaurant, the woman pulls out her mobile phone, clicks a button on the side, and speaks her thoughts about the meal. Her location and the time of day are recorded automatically along with her speech. The following morning, the woman goes to a website, sees the restaurant displayed on a map of Toronto, along with a summary and automatically determined ratings from her audio review, integrated with additional information about the restaurant such as address, etc. She looks it over, makes a small addition, and then shares it with her friends.

The scenario above illustrates an envisioned system that uses speech for *content creation*. In this paper, we describe our preliminary experiments in making such a system a reality. These experiments encourage us to believe that the underlying technology is sufficiently mature to use speech in this way. Mobile devices are uniquely suited to the problem, since both GPS coordinates and time-stamps can be associated with reviews to add valuable information for both processing and displaying information.

Behind the scenes, speech is uploaded to a cloud-based system. With a combination of automatic speech recognition (ASR) and natural language processing (NLP) technologies, key information is extracted from the dictated review. Words and phrases capturing features such as *food quality* or *service* are processed to automatically populate a form with extracted summary content and feature ratings.

One of the most important aspects of this scenario is that the user is in charge of the interaction the entire time. Users can describe an experience while it is fresh in their memory through an interface that is always available. When they have the time and the inclination, they can examine, review, and, ultimately, share it.

Several characteristics of this form of speech for *content creation* make it attractive from a technological perspective:

- It does not have to be real-time. As our scenario illustrates, the user simply speaks to a mobile device to enter content. Any further interaction takes place at the convenience of the user.
- It does not require a detailed word-by-word analysis of the input. Further processing of the text can be done using just recognized keywords/phrases in the user's input.
- It can be designed with multiple fallback mechanisms, such that any step of the process can be perceived as useful and beneficial to the user.

In this paper, we begin by describing previous work in ASR and NLP that helped inform the idea of using speech in this way. In section 3, we describe a data collection effort that gave us an initial corpus of audio reviews in the restaurant domain. Section 4 describes experiments using these data, with the goal of mining audio review data for useful information. Although both the data and experiments described in this paper are preliminary, we feel they are encouraging and provide motivation for using speech in this way.

2. PREVIOUS WORK

Previous work on large vocabulary ASR has shown that extracting meaning from spontaneous speech is possible, and

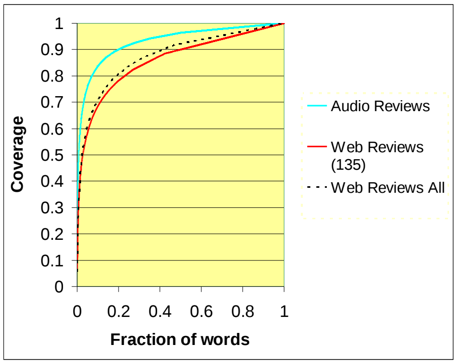

Fig. 1. Vocabulary growth as a function of word coverage. The top line represents the spoken reviews, the middle, dashed line is for the subset of Web-based reviews corresponding in size to the audio review data, and the bottom line is for the entire set of Web-based reviews.

does not necessarily involve a complete analysis of the input utterance. There is value in simply extracting keywords and phrases from speech data. The more sophisticated analysis needed for sentiment detection, although thus far only applied to text data, also involves processing only parts of the text.

Keyword/phrase spotting has long been used to perform *partial understanding* of spontaneous speech in systems where interaction is restricted to a single turn. The *How May I Help You* (*HMIHY*) system at AT&T [1] was one of the first to show the viability of extracting salient phrases from unconstrained speech input to perform call routing. Text categorization technology has also been applied to the same class of data, i.e., large-vocabulary, spontaneous speech [2]. The *What You See Is Almost What You Hear* principle for designing UIs [3] took this paradigm further to allow for searching and browsing ASR results in the voicemail domain.

Output of large vocabulary ASR has also been used to *segment* and *index* audio recordings. In developing the SpeechBot system, van Thong *et al.* found that information retrieval performance using text derived via ASR was actually higher than would be expected given the error rate of the ASR engine [4]. Suzuki *et al.* found they were able to perform keyword extraction on radio news using ASR and use that information to identify domains of individual segments [5].

Lecture browsing is another domain in which the output of ASR engines has been sufficiently accurate to provide real value [6, 7]. To provide entree into hours of otherwise unsegmented spontaneous speech, topics must be discovered and delimited automatically, using keywords and phrases. One of the models used to partition these data, the minimum-cut segmentation model, was originally developed on text data and has been subsequently found to be robust to recognition errors. In comparisons of performance using this model, only a moderate degradation in classification accuracy was observed for speech transcription vs. text [8].

Protocol A
1. What is the name of the restaurant?
2. Where is this restaurant located?
3. What type of cuisine does it serve?
4. What is its phone number?
5. Rate this restaurant on a scale of 1-5, where 1 is poor and 5 is excellent.
6. Rate the food quality on a scale of 1-5, where 1 is poor and 5 is excellent.
7. Rate the quality of service on a scale of 1-5, where 1 is poor and 5 is excellent.
8. Rate the atmosphere on a scale of 1-5, where 1 is poor and 5 is excellent.
9. Please review the restaurant and your experience there in your own words.

Table 1. Protocol A used for the ongoing data collection effort in spoken restaurant reviews.

Opinion mining and *sentiment detection* represent an extension of the paradigm of performing useful NLP in the absence of a complete parse of the input. These technologies work from tokenized collections of individual words. Many techniques employed for opinion mining, sentiment detection, and feature extraction utilize words and phrases found within the text rather than a complete analysis of each word as it functions within a sentence. Initial work in sentiment detection focussed on extracting the polarity of a given text, applied to written reviews for a variety of consumer products, as well as entities such as movies and restaurants [9, 10]. As the technology matured, it became possible to determine a more fine-grained rating, indicating a scale of sentiment [11, 12].

For completely unstructured text, such as that found in user-generated content on the Web, it is also useful to automatically extract information about individual features mentioned in reviews. Features are combined with gradient rating of attributes to enable even deeper insight into consumer-generated media [13, 14, 15, 16], It becomes possible to approach the insights of guides such as Zagat's or Consumer Reports, derived from a broad spectrum of opinions, with correspondingly little effort or time.

We feel that partial understanding is especially valuable for "one-shot" type applications, where there is no need for a more detailed analysis that will drive an ongoing dialogue. If an immediate and detailed interaction is not required, or even desired, the system has the luxury of performing computationally expensive processing while, at the same time, not having to completely understand everything the user spoke or involve them in a tedious confirmation dialogue.

Some techniques for sentiment detection make use of individual utterances and, therefore, utterance boundaries, for either computing the overall polarity of a given text [17] or to subset a larger text into just segments of interest for a market department, for example [18]. Although not all users will speak flawlessly complete sentences, we expect an underlying prosodic and language model to be present nonetheless. Automatic addition of periods and other punctuation has already been shown to be possible in speech and beneficial to performance of automatic speech recognizers [19, 20]. It has been further shown to help in identifying names in speech [21]. We examine the use of automatic boundary assignment in speech to aid in the extraction of keyphrases for describing restaurants in Section 4.1.

3. DATA COLLECTION

We collected an initial corpus of audio restaurant reviews in a laboratory setting, with subjects recruited from among a self-reported population who frequently eat out at restaurants and who were familiar with on-line restaurant review sites. Each of these subjects spoke to a Nokia E-72 handset instrumented for the purpose of data collection. Users saw the questions shown in Tables 1 and 2 on the handset and responded by clicking and holding to talk. All utterances were transcribed.

Subjects were randomly assigned to answer one of two questionnaires, both in the restaurant domain. In both questionnaires, the users were asked to rate the *food quality*, *service*, and *atmosphere* of each individual restaurant on a scale of 1-5, with 1 being least favorable and 5 most favorable. In one set of questions, which we call *Protocol A*, shown in Table 1, users were asked to assign a scalar value to each of the three attributes and then rate the restaurant and their experience as a whole in a single response. In the second set of questions, which we call *Protocol B*, shown in Table 2, users were asked to assign a scalar value *and* verbally describe each individual attribute. These users were also asked to provide an overall spoken review, in which they could repeat information previously spoken.

This data collection effort was designed to give us flexibility in designing an initial application and also to provide insight into review data elicited under slightly different protocols. Both sets of questionnaires were designed primarily to collect data that can be used to associate users' spoken reviews with an automatically derived value representing their sentiment about the restaurant and its attributes (described in Section 4). The first set of questions represents an ideal situation, i.e., where a user simply speaks in a free-form manner and we determine both features and polarity ratings. The latter set of data has been designed to capture specific information that might be useful to bootstrap training algorithms for automatically detecting specific features and assigning a graded sentiment representation for each. Both sets will be used for language model training.

Protocol B
1. What is the name of the restaurant?
2. Where is this restaurant located?
3. What type of cuisine does it serve?
4. What is its phone number?
5. Rate this restaurant on a scale of 1-5, where 1 is poor and 5 is excellent.
6. Rate the food quality on a scale of 1-5, where 1 is poor and 5 is excellent.
7. In words, please summarize the food quality.
8. Rate the quality of service on a scale of 1-5, where 1 is poor and 5 is excellent.
9. In words, please summarize the service.
10. Rate the atmosphere on a scale of 1-5, where 1 is poor and 5 is excellent.
11. In words, please summarize the atmosphere.
11. Please review the restaurant and your experience. Repeating information is okay.

Table 2. Protocol B used for the ongoing data collection.

We collected 135 reviews altogether, of which 73 had the additional audio reviews of individual restaurant features. Table 3 shows statistics on vocabulary size for various sets of review data. At the top of the table are numbers for the audio reviews collected using the protocol described above. Numbers are grouped by attribute described in the review, with "overall" used for reviews of the entire experience. The "overall" reviews are further broken down into those that were gathered under Protocol A and those under Protocol B. The average number of words per review is smaller for the individual feature reviews than for the overall reviews. We were surprised that Protocol B resulted in a larger average number of words for the overall reviews. The preceding questions on individual features may have "primed" users in some sense to have more to say. However, with such a small dataset it is difficult to draw conclusions.

The bottom of Table 3 contains statistics on vocabulary from text reviews, for both a large set (1640) of reviews and a smaller set, chosen at random and equivalent in size to the number of overall audio reviews. Surprisingly, the average number of words is smaller on the large set of text reviews than on the spoken reviews from the data collection effort. This set includes a number of very terse reviews, e.g., "great food, great service", a type not seen at all in the audio reviews, even for the individual attributes. Although we hesitate to draw conclusions from this preliminary effort, it seems speech encourages people to be verbose in their opinions.

Figure 1 charts vocabulary growth for three sets of data, the consolidated set of overall reviews from both audio protocols, compared with a similarly sized set of text reviews, and a much larger set of text reviews. Singleton vocabulary items on the right side of the curve include place and restaurant names (e.g., "paducah", "middletown"), unusual dishes (e.g., "fagioli") and descriptive adjectives not commonly used to describe food (e.g., "electrifying").

It is striking that the spoken review curve rises much more sharply at the beginning, and the slope of the linear growth trajectory at high vocabularies is much shallower, i.e., new words are introduced much less frequently. We hypothesize that people don't use rare words as often in speaking as in writing, perhaps due to cognitive load. However, another potentially even larger factor is the issue of misspelled words, as well as devices such as emoticons, mentions of urls, and creative punctuation. It is appealing that, while speech faces the issue of recognition errors, misspellings and other extraneous

Answer type	# reviews	Total words	# uniq. words	Avg. # words
Food quality	73	3222	792	44.1
Service	73	3450	734	47.2
Atmosphere	73	3924	881	53.6
Overall A	73	6210	1218	85.1
Overall B	62	5007	1126	80.7
Overall All	135	11217	1754	83.1
Text reviews (large)	1640	127606	9239	77.8
Text reviews (small)	135	12348	2457	91.5

Table 3. Statistics from data collection using the two different protocols.

textual material are no longer a factor.

4. APPLYING TEXT-BASED TECHNIQUES TO SPEECH

This section describes two preliminary experiments based on previously developed algorithms for key phrase extraction [15] and sentiment detection [13]. We are interested to see if text-based strategies for summarization and sentiment detection can be applied to audio data. We have thus far restricted our experiments mostly to the transcripts, although we begin to address speech-related issues through experiments on automatic sentence boundary detection in parse-based phrase extraction, and on sentiment detection on ASR output for a portion of the data.

4.1. Applying Phrase Extraction

In this section, we describe a set of experiments to extract key descriptive phrases from the speech transcripts, using NLP techniques involving a parse-and-paraphrase paradigm [15, 22] Our approach extracts key adjective-noun pairs guided by syntactic structure. For example, "The caesar salad I ordered was delicious" ⟶ "delicious caesar salad." We envision that these phrases could be presented as a succinct review summary in a Web browser, and could be linked to associated audio clips.

Our generic syntax-based grammar requires explicit lexical knowledge of all terminal words except for the major word classes, *noun* and *adjective*. For these open categories, it relies on a pre-assigned syntactic label, with an associated score for each label provided by a simple maximum entropy-based classifier. In our experiments, each word is assigned a score for *adjective*, *noun*, or *other*, based on six basic features, which do not depend on sentence context: three are word unigram frequencies and the other three are letter sequences within the word. We utilize a training set of over 200,000 utterances extracted from consumer-provided Web reviews to provide statistics on both word frequencies and letter trigram frequencies within the three classes.

	Percent Words Parse	# Phrases Extracted	Ave Words per 'Sentence'
manual	38.7%	259	16.6
autopunct	33.5%	233	17.7
integrated	76.6%	450	7.7

Table 4. Results of parse experiments on three configurations of the data. *Autopunct* involves a pre-assignment of sentence boundaries prior to parsing, whereas *integrated* assigns sentence boundaries based on maximizing parse coverage.

Parsing involves initially mapping words to preterminal categories via a finite state transducer (FST), which incorporates the label scores from the classifier [22]. The FST encodes as well a class n-gram language model, where the classes are the preterminal categories of the parse tree. The parser can then very efficiently parse the sentence, guided by an N-best list of preterminal sequences, which provide strong constraint to the parse search space. Selected open class preterminal categories support any words labeled *noun* or *adj* as terminal nodes. Nouns and their modifying adjectives are extracted based on parse constraints and filtered to reflect a subset of constructed descriptive phrases that contain words clearly associated with one of the major classes, *food*, *service*, *ambiance*, and *price* using techniques described in [15].

We utilized these techniques on three closely related data sets. The first one (*manual*) is the transcript of the audio recordings, manually pre-segmented into individual sentences. The second one (*autopunct*) utilized an n-gram language model (trained on Web-based review data) to automatically predict sentence breaks in the audio recordings. The third data set (*integrated*) utilized the entire transcript of each audio clip as a single input string, and the task of segmenting into sentences was integrated with parsing.

The monologue speech data are problematic for parsing, as they contain many disfluencies and agrammatical constructs, such as "the wait the waiter the waitress i should say it it's one the room is one run by one waitress and it's a small room and they are very friendly there." The pre-segmented sentences, which were on average 16.6 words long, yielded only 56% parse coverage, and these were of course mainly the shorter sentences.

Results are presented in Table 4. *Autopunct* yields both slightly longer sentences and fewer phrases than *manual*, likely due to the fact that sentence boundaries are at odd places and this affects parse coverage. The *integrated* yielded much shorter sentences, along with many more extracted phrases. Not only was it able to optimize boundaries for parsable segments, but it was also able to double as a robust-parsing strategy. It recovered valuable information from fragments extracted from the originally unparsable sentences. The input graph supported sentence breaks at every word, and an N-best list (N=40) of hypothesized preterminal sequences, including hypothesized sentence breaks, for the transcript of each unsegmented audio segment was produced by the FST. The parser optimized by choosing the hypothesis that maximized Nwords -Nsents, where Nwords is the total number of words contained in parsable segments, and Nsents is the total number of hypothesized sentences. Hence it favors longer sentences that are parsable, since we want to try to avoid separating an adjective from its associated noun, as in "I ordered the soup <break> which was delicious." It more than doubled the total number of "sentences," but was as a result able to recover valuable information from otherwise intractable sentences. Although we have not yet done a formal evaluation of the quality of the phrases, it is noteworthy that the *integrated* approach discarded fewer than 25% of the words compared with 66.5% for *autopunct*). As a consequence, it extracted substantially more phrases.

4.2. Applying Sentiment Detection

We used the Good Grief algorithm [13], which has been used on restaurant review data, as well as other forms of consumer-generated media, to perform state-of-the-art analysis of multiple related opinions. It produces a set of numeric scores, one for each aspect seen in the training data. For restaurant reviews, this includes features associated with restaurant reviews such as *food_quality*, *service*, and *atmosphere*, as well as a numeric rating for overall experience. The algorithm works by analyzing meta-relations about opinions, such as agreement and contrast, producing a joint model that is more expressive than individual ranking models.

The algorithm is evaluated using *rank loss* [23], the average distance between the true rank and the predicted rank. Because we did not have individual ratings from every user due to the data collection protocol, and because our dataset for audio reviews is already quite small, we test here on only the "overall" rating given by users.

We obtained manual transcripts for all of the audio data, and these were used both to evaluate Good Grief's performance on "perfect" recognition output, and to provide language training data for the recognition experiments, which were performed on a small subset of the data, set aside for that purpose. Thus, our very limited pilot ASR experiments involved ASR hypotheses created for just 25 audio reviews. We used the transcripts of the remaining 110 reviews to interpolate into an existing general purpose language model used by the ASR engine.

The results of these experiments are shown in Table 5. In each condition in the table, results are compared against a majority baseline. In general, Good Grief does well. Performance improves over the baseline, for both the transcript and the ASR outputs. Performance on the transcribed audio review is competitive with that on web-based data, as reported in [13].

While the transcripts of the subset reserved for the ASR experiments performed substantially worse than the full set, the ASR outputs performed substantially better than the transcripts. While we hesitate to draw firm conclusions from such a small dataset, we hypothesize that words that are critical to Good Grief performance are found reliably enough to make speech input a potentially viable option for sentiment detection algorithms.

	Rank loss	Rank loss baseline	# train reviews	# test reviews
Transcripts	0.60	0.89	1609	110
Transcripts	0.88	0.96	1609	25
ASR	0.68	0.96	1609	25

Table 5. Good Grief performance on audio transcripts versus speech recognition outputs. The top row shows results for all the transcripts. The second row shows results for the subset that were used for the ASR experiment. The third row gives the results for the ASR outputs on the same subset.

% found by ASR	Avg. wgt.	Examples
> 90%	6.30	exceptional, spectacular, pricey
< 90%	-1.19	there's, time, where, has

Table 6. Analysis of Good Grief weights for words above and below 90% recognition rate

To understand better why ASR output performed so well, we investigated the relationship between the weights Good Grief assigns to words and recognition performance on those words. Ideally, we would hope that words with high weights are well recognized. To roughly measure this aspect, we divided all words into two classes based on whether they were recognized over 90% of the time. The results are shown in Table 6, with example words for each category. Reassuringly, words that are recognized more reliably also have a higher average weight. The words with better recognition performance are, not surprisingly, often multi-syllabic and, in this domain, indicative of sentiment. We maintain that the bag-of-words approach of Good Grief is especially well suited to ASR outputs, as it does not depend on correct syntactic structure.

5. CONCLUSION AND FUTURE WORK

In this paper, we have described preliminary experiments using speech to create and annotate restaurant reviews. We provided some summary statistics, and reported results on two preliminary experiments related to review summarization and ranking. Our results are encouraging, and we look forward to continuing with this work. For phrase extraction, we plan to explore integrating the classifier with the output of a recognizer word graph. We also plan to collect more data, for both improving models of sentiment detection and for exploring issues in interface design on a mobile device.

6. ACKNOWLEDGEMENTS

The authors wish to thank Nokia Research Center for funding the data collection effort, and Mark Adler, Imre Kiss and Janet Slifka for their contributions to the work described here.

7. REFERENCES

[1] A. L. Gorin, B. A. Parker, R. M. Sachs, and J. G. Wilpon, “How May I Help You?,” *Speech Communication*, vol. 23, pp. 113–127, 1997.

[2] Robert E. Schapire and Yoram Singer, “BoosTexter: A boosting-based system for text categorization,” in *Machine Learning*, 2000, pp. 135–168.

[3] Steve Whittaker, Julia Hirschberg, Brian Amento, Litza Stark, Michiel Bacchiani, Philip Isenhour, Larry Stead, Gary Zamchick, and Aaron Rosenberg, “Scanmail: a voicemail interface that makes speech browsable, readable and searchable,” in *in Proceedings of CHI2002 Conference on Human Computer Interaction*. 2002, pp. 275–282, ACM Press.

[4] Jean-Manuel Van Thong, David Goddeau, Anna Litvinova, Beth Logan, Pedro Moreno, and Michael Swain, “Speechbot: a speech recognition based audio indexing system for the web,” in *Proc. of the 6th RIAO Conference*, 2000, pp. 106–115.

[5] Yoshimi Suzuki, Fumiyo Fukumoto, and Yoshihiro Sekiguchi, “Keyword extraction using term-domain interdependence for dictation of radio news,” in *Proceedings of the 17th international conference on Computational linguistics*, Morristown, NJ, USA, 1998, pp. 1272–1276, Association for Computational Linguistics.

[6] James Glass, Timothy Hazen, Scott Cyphers, Igor Malioutov, David Huynh, and Regina Barzilay, “Recent progress in the MIT spoken lecture processing project,” in *Proc., Interspeech*, 2007.

[7] Cosmin Munteanu, Gerald Penn, and Xiaodan Zhu, “Improving automatic speech recognition for lectures through transformation-based rules learned from minimal data,” in *ACL-IJCNLP ’09: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*, Morristown, NJ, USA, 2009, pp. 764–772, Association for Computational Linguistics.

[8] Igor Malioutov and Regina Barzilay, “Minimum cut model for spoken lecture segmentation,” in *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, Morristown, NJ, USA, 2006, pp. 25–32, Association for Computational Linguistics.

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up?: sentiment classification using machine learning techniques,” in *EMNLP ’02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Morristown, NJ, USA, 2002, pp. 79–86, Association for Computational Linguistics.

[10] Kamal Nigam and Matthew Hurst, “Towards a robust metric of opinion,” in *Proc., AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 2004.

[11] Kushal Dave, Steve Lawrence, and David M. Pennock, “Mining the peanut gallery: opinion extraction and semantic classification of product reviews,” in *WWW ’03: Proceedings of the 12th international conference on World Wide Web*, New York, NY, USA, 2003, pp. 519–528, ACM.

[12] Andrew B. Goldberg and Xiaojin Zhu, “Seeing stars when there aren’t many stars: graph-based semi-supervised learning for sentiment categorization,” in *TextGraphs ’06: Proceedings of TextGraphs: the First Workshop on Graph Based Methods for Natural Language Processing on the First Workshop on Graph Based Methods for Natural Language Processing*, Morristown, NJ, USA, 2006, pp. 45–52, Association for Computational Linguistics.

[13] Benjamin Snyder and Regina Barzilay, “Multiple aspect ranking using the Good Grief algorithm,” in *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, 2007, pp. 300–307.

[14] Ivan Titov and Ryan McDonald, “Modeling online reviews with multi-grain topic models,” in *WWW ’08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, 2008, pp. 111–120, ACM.

[15] Jingjing Liu and Stephanie Seneff, “Review sentiment scoring via a parse-and-paraphrase paradigm,” in *EMNLP ’09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, 2009, pp. 161–169, Association for Computational Linguistics.

[16] Narendra Gupta, Giuseppe di Fabbrizio, and Patrick Haffner, “Capturing the stars: Predicting rankings for service and product reviews,” in *Proc., NAACL HLT 2010 Workshop on Semantic Search*, 2010.

[17] Bo Pang and Lillian Lee, “Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales,” in *ACL ’05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Morristown, NJ, USA, 2005, pp. 115–124, Association for Computational Linguistics.

[18] Matthew Hurst and Kamal Nigam, “Retrieving topical sentiments from online document collections,” in *In Document Recognition and Retrieval XI*, 2004, pp. 27–34.

[19] Yang Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, “Enriching speech recognition with automatic detection of sentence boundaries and disfluencies,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[20] Kolár̂, Ŝvec Jáchym, and Josef Psutka, “Automatic punctuation annotation in czech broadcast news speech,” in *Proc., SPECOM*, 2004.

[21] D. Hillard, Z. Huang, H. Ji, R. Grishman, D. Hakkani-tur, M. Harper, M. Ostendorf, and W. Wang, “Impact of automatic comma prediction on pos/name tagging of speech,” in *Proc. of the IEEE/ACL 2006 Workshop on Spoken Language Technology*, 2006, pp. 58–61.

[22] Yushi Xu, Stephanie Seneff, Alice Li, and Joseph Polifroni, “Semantic understanding by combining extended CFG parser with HMM model,” in *submitted to these proceedings*, 2010.

[23] Koby Crammer and Yoram Singer, “Pranking with ranking,” in *Advances in Neural Information Processing Systems 14*. 2001, pp. 641–647, MIT Press.