An Unsupervised Aspect-Sentiment Model for Online Reviews

Samuel Brody
Dept. of Biomedical Informatics
Columbia University
samuel.brody@dbmi.columbia.edu

Noemie Elhadad
Dept. of Biomedical Informatics
Columbia University
noemie@dbmi.columbia.edu

Abstract

With the increase in popularity of online review sites comes a corresponding need for tools capable of extracting the information most important to the user from the plain text data. Due to the diversity in products and services being reviewed, supervised methods are often not practical. We present an unsupervised system for extracting aspects and determining sentiment in review text. The method is simple and flexible with regard to domain and language, and takes into account the influence of aspect on sentiment polarity, an issue largely ignored in previous literature. We demonstrate its effectiveness on both component tasks, where it achieves similar results to more complex semi-supervised methods that are restricted by their reliance on manual annotation and extensive knowledge sources.

1 Introduction

Online review sites continue to grow in popularity as more people seek the advice of fellow users regarding services and products. Unfortunately, users are often forced to wade through large quantities of written data in order to find the information they want. This has led to an increase in research in the areas of opinion mining and sentiment analysis, with the aim of providing systems that can automatically analyze user reviews and extract the information most relevant to the user.

One example of such an application is generating a summary of the important factors mentioned in the reviews of a product (see Lerman et al. 2009). Another application is comparing two similar products. In this case, it is important to present to the user the aspects in which the products differ, rather than just provide a general star rating. A third example is systems for generating automatic recommendations, based on similarity between products, user reviews, and history of previous purchases. These types of application require an underlying framework to identify the important aspects of the product (also known as *features* or *attributes*), and the sentiment expressed by the review writer.

Unsupervised Methods are desirable for this task, for two reasons. First, due to the wide range and variety of products and services being reviewed, the framework must be robust and easily transferable between domains. The second reason is the nature of the data. Online reviews are often short and unstructured, and may contain many spelling and grammatical errors, as well as slang or specialized jargon. These factors often present a problem to methods relying exclusively on dictionaries, manually-constructed knowledge resources, and gazetteers, as they may miss out on an important aspect of the product or an indicator of sentiment. Unsupervised methods, on the other hand, are not influenced by the lexical form, and can handle unknown words or word-forms, provided they occur frequently enough. This insures that any emergent topic that is salient in the data will be addressed by the system.

In this paper, we present an unsupervised system which addresses the core tasks necessary to enable advanced applications to handle review data. We introduce a local topic model, which works at the sentence level and employs a small number of topics, to automatically infer the aspects. For sentiment detection, we present a method for automatically deriving an unsupervised seed set of positive and negative adjectives that replaces the manually constructed ones commonly used in the literature. Our approach is specifically designed to take into account the inter-

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 804–812,
Los Angeles, California, June 2010.

action between the two tasks.

The rest of the paper is structured as follows. In Sec. 2 we provide relevant background, and place our method in the context of previous work in the field. We describe the data we used in Sec. 3, and our experiments on the aspect and sentiment-polarity components in Sec. 4 and 5, respectively. We conclude in Sec. 6 with a discussion of our results and findings and directions for future research.

2 Previous Approaches

In this paper, we focus on the detection of two principle elements in the review text: aspects and sentiment. In previous work these elements have been treated, for the most part, as two separate tasks.

Aspect The earliest attempts at aspect detection were based on the classic information extraction (IE) approach of using frequently occurring noun phrases (e.g., Hu and Liu 2004). Such approaches work well in detecting aspects that are strongly associated with a single noun, but are less useful when aspects encompass many low frequency terms (e.g., the *food* aspect of restaurants, which involves many different dishes), or are abstract (e.g. *ambiance* can be described without using any concrete nouns at all). Common solutions to this problem involve clustering with the help of knowledge-rich methods, involving manually-constructed rules, semantic hierarchies, or both (e.g., Popescu and Etzioni 2005, Fahrni and Klenner 2008). Titov and McDonald (2008b) underline the need for unsupervised methods for aspect detection. However, according to the authors, existing topic models, such as standard Latent Dirichlet Allocation (LDA) (Blei et al., 2003), are not suited to the task of aspect detection in reviews, because they tend to capture global topics in the data, rather than rateable aspects pertinent to the review. To address this problem, they construct a multi-grain topic model (MG-LDA), which attempts to capture two layers of topics - global and local, where the local topics correspond to rateable aspects. MG-LDA distinguishes tens of local topics, but the many-to-one mapping between these and rateable aspects is not explicit in the system. To resolve this issue, the authors extend their model in Titov and McDonald (2008a) and attempt to infer such a mapping with the help of aspect-specific ratings provided along with the review text.

Sentiment Sentiment analysis has been the focus of much previous research. In this discussion, we will only mention work directly related to our own. For a comprehensive survey of the subject, the reader is directed to Pang and Lee (2008).

Most previous approaches rely on a manually constructed lexicon of terms which are strongly positive or negative regardless of context. This information on its own is usually insufficient, due to lack of coverage and the fact that sentiment is often expressed through words whose polarity is highly domain and context specific. If a sentiment lexicon is available for one domain, domain adaptation can be used, provided the domains are sufficiently similar (Blitzer et al., 2007). Another common solution is through bootstrapping - using a seed group of terms with known polarity to infer the polarity of domain specific terms (e.g., Fahrni and Klenner 2008; Jijkoun and Hofmann 2009). The most minimalist example of this approach is Turney (2002), who used only a single pair of adjectives (*good* and *poor*) to determine the polarity of other terms through mutual information. For Chinese, Zagibalov and Carroll (2008) use a single seed word meaning *good*, and six common indicators of negation in their bootstrapping approach. Often, when using a context independent seed, large amounts of domain-specific data are required, in order to obtain sufficient co-occurrence statistics. Commonly, web queries are used to obtain such data.

Independently of any specific task, Hatzivassiloglou and McKeown (1997) present a completely unsupervised method for determining the polarity of adjectives in a large corpus. A graph is created, in which adjectives are nodes, and edges between them are weighted according to a (dis)similarity function based primarily on whether the two adjectives occurred in a conjunction or disjunction in the corpus. A heuristic approach is then used to split the graph in two. The group containing the adjectives with the higher average frequency is labeled as positive, and the other as negative.

Combined Approaches Aspects can influence sentiment polarity within a single domain. For example, in the restaurant domain, *cheap* is usually positive when discussing food, but negative when discussing the decor or ambiance. Many otherwise neutral terms (e.g., *warm, heavy, soft*) acquire a sentiment polarity in the context of a specific aspect.

Recent work has addressed this interaction in different ways. Mei et al. (2007) present a form of domain adaptation using an LDA model which treats positive and negative sentiment as two additional topics. Fahrni and Klenner (2008) directly address the specificity of sentiment to the word it is modifying. Aspects are defined by a manually specified subset of the Wikipedia category hierarchy. For sentiment, the authors use a seed set of positive and negative adjectives, and iteratively propagate sentiment polarity through conjunction relations (like those used by Hatzivassiloglou and McKeown 1997, above). Web queries are used to overcome the sparsity issue of these highly-specific patterns. In the IE setting, Popescu and Etzioni (2005) extract frequent terms, and cluster them into aspects. The sentiment detection task is formulated as a Relaxation Labeling problem of finding the most likely sentiment labels for opinion-bearing terms, while satisfying as many local constraints as possible. The authors use a variety of knowledge sources, web queries, and hand crafted rules to detect relations between terms (e.g., meronymy). These relations are used both for the clustering, and as a basis for the constraints.

Our approach is designed to be as unsupervised and knowledge-lean as possible, so as to make it transferable across different types of products and services, as well as across languages. Aspects are determined via a local version of LDA, which operates on sentences, rather than documents, and employs a small number of topics that correspond directly to aspects. This approach overcomes the problems of frequent-term methods, as well as the issues raised by Titov and McDonald (2008b). We use morphological negation indicators to automatically create a seed set of highly relevant positive and negative adjectives, which are guaranteed to be pertinent to the aspect at hand. These automatically-derived seed sets achieve comparable results to the use of manual ones, and the work of Zagibalov and Carroll (2008) suggests that the use of negation can be easily transfered to other languages.

3 Data

Our primary dataset is the publicly available corpus used in Ganu et al. (2009). It contains over 50,000 restaurant reviews from Citysearch New York[1]. Additionally, to demonstrate the domain independence of our system, we collected 1086 reviews for four leading netbook computers from Amazon.com.

For evaluation purposes, we used the annotated dataset from Ganu et al. (2009), which is a subset of 3,400 sentences from the Citysearch corpus. These sentences were manually labeled for aspect and sentiment. There were six manually defined aspect labels - *Food & Drink*, *Service*, *Price*, *Atmosphere*, *Anecdotes* and *Miscellaneous*. A sentence could contain multiple aspects, but, for our evaluation, we used only sentences with a single label. For sentiment, each sentence was given a single value - *Positive, Negative, Neutral* or *Conflict* (indicating a mixture of positive and negative sentiment).

We were also provided with a seed set of 128 positive and 88 negative adjectives used by Fahrni and Klenner (2008), which were specifically selected to be domain and target independent.

For the purpose of the experiments presented here, we focused on sentences containing noun-adjective pairs. Such pairs are one of the most common way of expressing sentiment about an aspect and allow us to capture the interaction between the two.

4 Aspect

4.1 Methodology

In order to infer the salient aspects in the data, we employed the following steps:

Local LDA We used a standard implementation[2] of LDA. In order to prevent the inference of global topics and direct the model towards rateable aspects (see Sec. 2), we treated each sentence as a separate document. The output of the model is a distribution over inferred aspects for each sentence in the data. The parameters we employed were standard, out-of-the-box settings ($\alpha = 0.1, \beta = 0.1$, 3000 iterations), with no specific tuning to our data. We ran the algorithm with the number of aspects ranging from 10 to 20, and employed a cluster validation scheme (see below) to determine the optimal number.

Model Order The issue of model order, i.e., determining the correct number of clusters, is an important element in unsupervised learning. A common

[1] http://newyork.citysearch.com/

[2] GibbsLDA++, by Xuan-Hieu Phan. Available at http://gibbslda.sourceforge.net/.

approach (Levine and Domany, 2001; Lange et al., 2004; Niu et al., 2007) is to use a cluster validation procedure. In such a procedure, different model orders are compared, and the one with the most consistent clustering is chosen. For the purpose of the validation procedure, we have a cluster corresponding to each aspect, and we label each sentence as belonging to the cluster of the most probable aspect.

Given the collection of sentences in our data, D, and two connectivity matrices C and $\hat{C}$, where a cell i,j contains 1 if sentences d_i and d_j belong to the same cluster, we define a consistency function F (following Niu et al. 2007):

$$F(C,\hat{C}) = \frac{\sum_{i,j} 1\{C_{i,j} = \hat{C}_{i,j} = 1, d_i, d_j \in \hat{D}\}}{\sum_{i,j} 1\{C_{i,j} = 1, d_i, d_j \in \hat{D}\}} \quad (1)$$

We then employ the following procedure:

1. Run the LDA model with k topics on D to obtain connectivity matrix C_k.
2. Create a comparison connectivity matrix R_k based on uniformly drawn random assignments of the instances.
3. Sample random subset D^i of size $\delta|D|$ from D.
4. Run the LDA model on D^i to obtain connectivity matrix C_k^i.
5. Create a comparison matrix R_k^i based on uniformly drawn random assignments of the instances in D^i.
6. Calculate $score_i(k) = F(\hat{C},C) - F(\hat{R},R)$ where F is given in Eq. 1.
7. Repeat steps 3 to 6 q times.
8. Return the average score over q iterations.

This procedure calculates the consistency of our clustering solution, using a similar sized random assignment for comparison. It does this on q subsets to reduce the effects of chance. The k with the highest score is chosen. In our experiments, we used $q = 5, \delta = 0.9$. For both our datasets (restaurants and netbooks), the highest-scoring k was 14.

Determining Representative Words For each aspect, we list all the nouns in the data according to a score based on their mutual information with regard to that aspect.

$$Score_a(w) = p(w,a) \cdot log\frac{p(w,a)}{p(w) \cdot p(a)} \quad (2)$$

Where $p(w), p(a), p(w,a)$ are the probabilities, according to the LDA model, of the word w, the aspect a, and the word w labeled with aspect a, respectively.

We then select, for each aspect, the top k_a ranking words, such that they cover 75% of the word-instances labeled by the LDA model with aspect label a. Due to the skewed frequency distribution of words, this is a relatively small portion of the words (typically 100-200). This set of representative words for each aspect is used in the sentiment component of our system (see Sec. 5.1).

4.2 Inferred Aspects

Table 1 presents the aspects inferred by our system for the restaurant domain. The inferred aspects cover all those defined in the manual annotation, but also distinguish between a finer granularity of aspects, based solely on the review text, e.g., between physical environment and ambiance, and between the attitude of the staff and the quality of the service.

In order to demonstrate that our method can be transfered between very different domains and categories of products, we also ran our algorithm on our set of netbook reviews. The inferred aspects are presented in Table 2. The system identifies important aspects relevant to our data. Some of these (e.g., software, hardware) might be suggested by human annotators, but some would probably be missed unless the annotators carefully read through all the reviews, e.g., the *Memory* aspect, which includes advice about upgrading specific models. This capability of our system is important, as it demonstrates that our method can be used to produce customized comparisons for the user and will take into account the important common factors, as well as the unique aspects of each item.

4.3 Evaluation

To determine the quality of our automatically inferred aspects, we compared the output of our system to the sentence-level manual annotation of Ganu et al. (2009). To each sentence in the data, the LDA model assigns a distribution $\{P(a)\}_{a \in A}$ over the set A of inferred aspects. By defining a threshold t_a for each aspect, we can label a sentence as belonging to aspect a if $P(a) > t_a$. By varying the threshold t_a we created precision-recall curves for the top three rateable aspects in the restaurant domain, shown in

Inferred Aspect	Representative Words	Manual Aspect
Main Dishes	chicken, sauce, rice, cheese, spicy, salad,	Food & Drink
Bakery	hot, delicious, dessert, bagels, bread, chocolate	
Food - General	menu, fresh, sushi, fish, chef, cuisine	
Wine & Drinks	wine, list, glass, drinks, beer, bottle	
Ambiance / Mood	great, atmosphere, wonderful, music, experience, relaxed	Atmosphere
Physical Atmosphere	bar, room, outside, seating, tables, cozy, loud	
Staff	service, staff, friendly, attentive, busy, slow	Staff
Service	table, order, wait, minutes, reservation, forgot	
Value	portions, quality, worth, size, cheap	Price
Anecdotes	dinner, night, group, friends, date, family	Anecdotes
Anecdotes	out, back, definitely, around, walk, block	
General	best, top, favorite, city, NYC	Misc.
Misc. - Location	never, restaurant, found, Paris, (New) York, location	
Misc.	place, eat, enjoy, big, often, stuff	

Table 1: List of automatically inferred aspects for the restaurant domain, with some representative words for each aspect (middle), and the corresponding aspect label from the manual annotation (right). Labels (left) were assigned by the authors.

Aspect	Representative Words	Aspect	Representative Words
Performance	power, performance, mode, fan, quiet	Mouse	mouse, right, touchpad, pad, buttons, left
Hardware	drive, wireless, bluetooth, usb, speakers, webcam	General	great, little, machine, price, netbook, happy
Memory	ram, 2GB, upgrade, extra, 1GB, speed	Purchase	amazon, purchased, bought, weeks, ordered
Software	using, office, software, installed, works, programs	Looks	looks, feel, white, finish, blue, solid, glossy
Usability	internet, video, web, movies, music, email, play	OS	windows, xp, system, boot, linux, vista, os
Portability	around, light, work, portable, weight, travel	Battery	battery, life, hours, time, cell, last
Comparison	netbooks, best, reviews, read, decided, research	Size	screen, keyboard, size, small, enough, big

Table 2: List of automatically inferred aspects for the netbook dataset, with representative words for each aspect .

Figure 1[3]. Although the data used in Titov and McDonald (2008a) was unavailable for direct comparison, our method exhibits similar behavior and performance (compare Fig. 4, there) on a domain with similar characteristics (abstract aspects which encompass many low frequency words). This demonstrates that our local version of LDA with few topics overcomes the issues which confronted the authors of that work (i.e., global topics and many-to-one mapping of topics to aspects), without requiring specially designed models or additional information in the form of user-provided aspect-specific ratings (see Sec. 2).

We believe the reason for this stems from the composition of online reviews. Since many reviews have similar mixtures of local topics (e.g., food, service), standard LDA prefers global topics, which distinguish more strongly between *reviews* (e.g., cuisine type, restaurant type). However, when employed at the sentence level, local topics (corresponding to rateable aspects) provide a stronger way to distinguish between individual *sentences*.

[3]We combined the probabilities of all the inferred aspects that match a single manually assigned aspect, according to the mapping in Table 1.

5 Sentiment

5.1 Methodology

For determining sentiment polarity, we developed the following procedure. For each aspect, we extracted the relevant adjectives, built a conjunction graph, automatically determined the seed set (or used a manual one, for comparison), and propagated the polarity scores to the rest of the adjectives. Details of each step are described below.

Extracting Adjectives As a pre-processing step, we parsed our data (using RASP, Briscoe and Carroll 2002). The parsed output was used to detect negation and conjunction. If an adjective *A* partic-

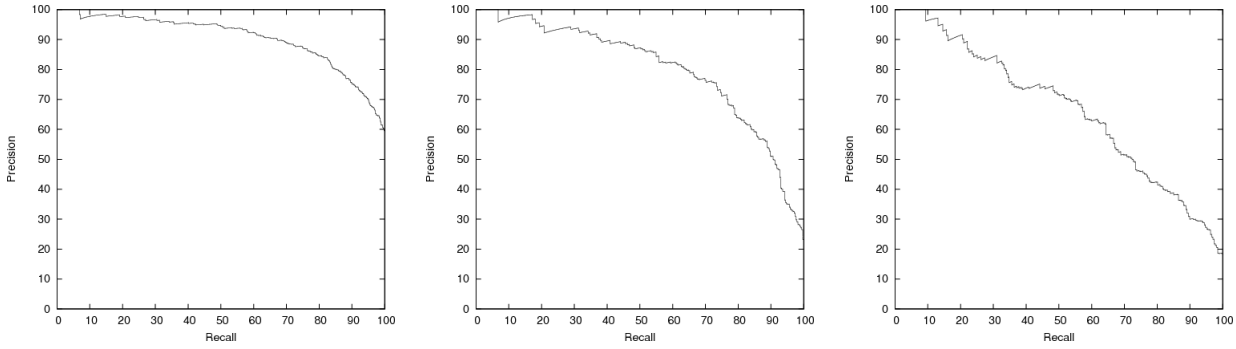

Figure 1: Precision / Recall curves for the top three rateable aspects: (a) *Food*, (b) *Service*, and (c) *Atmosphere*.

ipated in a negation in the sentence, it was replaced by a new adjective *not-A*. We then extract all cases where an adjective modified a noun. For example, from the sentence *"The food was tasty and hot, but our waiter was not friendly."* we can extract the pairs *(tasty, food), (hot, food), (not-friendly, waiter)*.

Building the Graph Our method for determining sentiment polarity is based on an adaptation of Hatzivassiloglou and McKeown (1997) (see Sec. 2).

Several issues confronted us when attempting to adapt their method to our task. In the original article, adjectives with no orientation were ignored. It is unclear how this can be easily done in an unsupervised fashion, and such sentiment-neutral adjectives are ubiquitous in real-world data. Furthermore, adjectives whose orientation depended on the context were also ignored. These are of particular interest in our task, and are likely to be missing or incorrectly labeled in standard sentiment dictionaries. For our purposes, since we need to handle adjectives expressing various shades of sentiment, not only strongly positive or negative ones, we are interested in a scoring method, rather than a binary labeling. Also, we do not want to use a general corpus, but rather the text from the reviews themselves. This usually means a much smaller corpus than the one used in the original paper, but has the advantage of being domain specific.

Our method of building the polarity graph differed in several ways from the original. First, we did not use disjunctions (e.g., 'but') as indicators of opposite polarity. The reason for this was that, in our domain of online reviews, disjunctions often did not convey contrast in polarity, but rather in perceived expectations, e.g., *"dainty but strong necklace"*, and *"cheap but delicious food"*.

Instead of using regular expressions to capture explicit conjunctions, we retrieved all cases where our parser indicated that two adjectives modified a single noun in the same sentence.

To ensure that aspect-specific adjectives are handled correctly, we built a separate graph for each aspect, by selecting the cases where the modified noun was one of the representative words for that aspect (see Sec. 4.1).

Constructing a Seed Set We used morphological information and explicit negation to find pairs of opposite polarity. Specifically, adjective pairs which were distinguished only by one of the prefixes *'un', 'in', 'dis', 'non',* or by the negation marker *'not-'* were selected for the seed set. Starting with the most frequent pair, we assigned a positive polarity to the more frequent member of the pair.

Then, in order of decreasing frequency, we assigned polarity to the other seed pairs, based on the shortest path either of the members had to a previously labeled adjective. That member received its neighbor's polarity, and the other member of the pair received the opposite polarity. When all pairs were labeled, we corrected for misclassifications by iterating through the pairs and reversing the polarity if that improved consistency, i.e., if it caused the members of the pair to match the polarities of more of their neighbors. Finally, we reverse the polarity of the seed groups if the negative group has a higher total frequency.

Propagating Polarity Our propagation method is based on the label propagation algorithm of Zhu and Ghahramani (2002). The adjectives in the positive and negative seed groups are assigned a polarity

score of 1 and 0, respectively. All the rest start with a score of 0.5. Then, an update step is repeated. In update iteration t, for each adjective x that is *not in the seed*, the following update rule is applied:

$$p^t(x) = \frac{\Sigma_{y \in N(x)} w(y,x) \cdot p^{t-1}(y)}{\Sigma_{y \in N(x)} w(y,x)} \quad (3)$$

Where $p^t(x)$ is the polarity of adjective x at step t, $N(x)$ is the set of the neighbors of x, and $w(y,x)$ is the weight of the edge connecting x and y. We set this weight to be $1 + log(\#mod(y,x))$ where $\#mod(y,x)$ is the number of times y and x both modified a single noun. The update step is repeated to convergence.

5.2 Aspect-Specific Gold Standard

To evaluate the performance of the sentiment component of our system, we created an aspect-specific gold standard. For each of the top eight automatically inferred aspects (corresponding to the *Food, Service* and *Atmosphere* aspects in the annotation), we constructed a polarity graph, as described in Sec. 5.1. We retrieved a list of all adjectives that participated in five or more modifications of nouns from that specific aspect). Table 3 lists the number of such adjectives in each aspect. We split the data into ten portions and, for each portion, asked two volunteers to rate each adjective according to the polarity of the sentiment it expresses *in the context of the specified aspect*. The judges could select from the following ratings: *Strongly Negative*, *Weakly Negative*, *Neutral*, *Weakly Positive*, *Strongly Positive*, and *N/A*. As expected, exact inter-annotator agreement was low - only 54%, but when considering two adjacent ratings as equivalent (i.e, Strongly vs. Weakly Negative or Positive, and Neutral vs. Weakly Negative or Positive), agreement was 93.3%. This indicates there is some difficulty distinguishing between the fine-grained categories we specified, but high agreement at a coarser level, which advocates using a ranking approach for evaluation (see also Pang and Lee 2005). We therefore translated the annotator ratings to a numerical scale, from -2 (Strongly Negative) to $+2$ (Strongly Positive) at unit intervals. After discarding adjectives where one or more annotators gave a 'N/A' tag, we averaged the two annotator numerical scores, and used this data as the gold standard for our evaluation.

Aspect	# Adj.	# Rated	% Neu.
Mood	293	206	17%
Staff	155	122	3%
Main Dishes	287	185	25%
Physical Atmo.	161	103	21%
Bakery	180	129	23%
Food - General	192	144	28%
Wine & Drinks	111	75	18%
Service	89	57	5%
Total	1468	1021	–

Table 3: For each aspect, the number of frequently occurring adjectives for each aspect (# Adj.), number of adjectives remaining after removing those labeled 'N/A' (# Rated), and percent of rated adjectives labeled 'Neutral' by both annotators (% Neu.).

	Auto.		Manual	
Aspect	τ_k	D_k	τ_k	D_k
Mood	0.53	0.23	0.56	0.22
Staff	0.57	0.22	0.60	0.20
Main Dishes	0.19	0.40	0.38	0.31
Physical Atmo.	0.34	0.33	0.25	0.37
Bakery	0.33	0.33	0.35	0.33
Food - General	0.19	0.41	0.41	0.30
Wine & Drinks	0.32	0.34	0.52	0.24
Service	0.41	0.30	0.54	0.23
Average	0.36	0.32	0.45	0.27

Table 4: Kendall coefficient and distance scores for eight inferred aspects.

5.3 Evaluation Measures

Kendall's tau coefficient (τ_k) and Kendall's distance (D_k) are commonly used (e.g., Jijkoun and Hofmann 2009) to compare rankings. These measures look at the number of pairs of ranked items that agree or disagree with the ordering in the gold standard. The value of τ_k ranges from -1 (perfect disagreement) to 1 (perfect agreement), with 0 indicating an almost random ranking. The value of D_k ranges from 0 (perfect agreement) to 1 (perfect disagreement). It is important to note that only pairs that are ordered in the gold standard are used in the comparison.

5.4 Evaluation Results

Table 4 reports Kendall's coefficient (τ_k) and distance (D_k) values for our method when using our automatically derived seed set (Auto.). For comparison, we ran our procedure using the manually compiled seed set (Manual) of Fahrni and Klenner

Food - General: Mexican, *French*, Eastern, Turkish, European, Tuscan, Mediterranean, American, Cuban, Thai, Peruvian, Spanish, Korean, Vietnamese, Indian, African, Japanese, Italian, *Chinese*, Asian
Mood: Vietnamese, Brazilian, Turkish, Eastern, Caribbean, Cuban, Italian, Spanish, Japanese, European, Mediterranean, Colombian, Mexican, Asian, Indian, Thai, British, American, *French*, Korean, *Chinese*, Russian, Moroccan
Staff: British, European, *Chinese*, Indian, American, Spanish, Asian, Italian, *French*

Table 5: Polarity ranking of cuisine adjectives (from most positive) for three aspects.

(2008). Using the manual seed set obtains results that correspond better to our gold standard. Our automatic method also achieves good results, and can be used when a manual seed set is not available. More importantly, correlation with the gold standard may not indicate better suitability to the sentiment detection task in reviews. For instance, it is interesting to note that the worst correlation scores were on the *Main Dishes* and *Food - General* aspects. If we compare to Table 3, we can see these aspects have the highest percentage of adjectives rated as neutral by the annotators. However, in many cases, these adjectives actually carry some sentiment in their context. An example of this are adjectives describing the type of cuisine, which are objective, and therefore usually considered neutral by annotators. Table 5 shows the automatic ranking of cuisine type from positive to negative in three aspects. It is interesting to see that the rankings change according to the aspect, and certain cuisines are strongly associated with specific aspects and not with others. This is supported by Ganu et al. (2009), who observed during the annotation that, in the restaurant corpus, French and Italian restaurants were strongly associated with the service aspect. This trend can be identified automatically by our method, and at a much more detailed level than that noticed by a human analyzing the data.

6 Discussion & Future Work

Our experiments confirm the value of a fully unsupervised approach to the tasks of aspect detection and sentiment analysis. The aspects are inferred from the data, and are more representative than manually derived ones. For instance, in our restaurant domain, the manually constructed aspect list omitted or over-generalized some important aspects, while over-representing others. There was no separate *Drinks* category, even though it was strongly present in the data. The *Service* aspect, dealing with waiting time, reservations, and mistaken orders, was an important emergent aspect on its own, but was grouped under *Staff* in the manual annotation.

Adjectives can convey different sentiments depending on the aspect being discussed. For example, the adjective *'warm'* was ranked very positive in the *Staff* aspect, but slightly negative in the *General Food* aspect. A knowledge-rich approach might ignore such adjectives, thereby missing important elements of the review.

Finally, as online reviews belong to an informal genre, with inventive spelling and specialized jargon, it may be insufficient, for both aspect and sentiment, to rely only on lexicons. For example, our restaurant reviews included spelling errors such as *desert, decour/decore, anti-pasta, creme-brule, sandwhich, omlette, exelent, tastey,* as well as at least six different common misspellings of *restaurant*. There were also specialized terms, such as *Korma, Edamame, Dosa* and *Pho*, all of which do not appear in common dictionaries, and creative use of adjectives, such as *orgasmic* and *New-Yorky*.

This work has opened many avenues for future research and improvements. So far, we focused on adjectives as sentiment indicators, however, there have been studies showing that other parts of speech can be very helpful for this task (e.g., Pang et al. 2002; Benamara et al. 2007). Also, it would be interesting to take a closer look at the interactions between aspect and sentiment, especially at a multiple-sentence level (see Snyder and Barzilay 2007). Finally, we feel that the true test of the usability of our system should be through an application, and intend to proceed in that direction.

Acknowledgments

We'd like to thank Angela Fahrni and Manfred Klenner for kindly allowing us access to their data and annotation. We also wish to thank the volunteer annotators. This work was partially supported by a Google Research Award.

References

Benamara, Farah, Carmine Cesarano, Antonio Picariello, Diego Reforgiato, and V. S. Subrahmanian. 2007. Sentiment analysis: Adjectives and adverbs are better than

adjectives alone. In *Proc. of the International Conference on Weblogs and Social Media (ICWSM)*.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*. ACL, Prague, Czech Republic, pages 440–447.

Briscoe, Ted and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proc. of the 3rd LREC*. Las Palmas, Gran Canaria, pages 1499–1504.

Fahrni, Angela and Manfred Klenner. 2008. Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives. In *Proc.of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention*. pages 60 – 63.

Ganu, Gayatree, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In *WebDB*.

Hatzivassiloglou, Vasileios and Kathleen R. McKeown. 1997. Predicting the semantic orientation of adjectives. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics*. ACL, Madrid, Spain, pages 174–181.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. In *KDD '04: Proc. of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, New York, NY, USA, pages 168–177.

Jijkoun, Valentin and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In *Proc. of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. ACL, Athens, Greece, pages 398–405.

Lange, Tilman, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. 2004. Stability-based validation of clustering solutions. *Neural Comput.* 16(6):1299–1323.

Lerman, Kevin, Sasha Blair-Goldensohn, and Ryan McDonald. 2009. Sentiment summarization: evaluating and learning user preferences. In *EACL '09: Proc. of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. ACL, Morristown, NJ, USA, pages 514–522.

Levine, Erel and Eytan Domany. 2001. Resampling method for unsupervised estimation of cluster validity. *Neural Comput.* 13(11):2573–2593.

Mei, Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai. 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *WWW '07: Proc. of the 16th international conference on World Wide Web*. ACM, New York, NY, USA, pages 171–180.

Niu, Zheng-Yu, Dong-Hong Ji, and Chew-Lim Tan. 2007. I2r: three systems for word sense discrimination, chinese word sense disambiguation, and english word sense disambiguation. In *SemEval '07: Proc. of the 4th International Workshop on Semantic Evaluations*. ACL, Morristown, NJ, USA, pages 177–182.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proc. of the ACL*. pages 115–124.

Pang, Bo and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1-2):1–135.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proc. of the conference on Empirical methods in natural language processing*. ACL, Morristown, NJ, USA, pages 79–86.

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting product features and opinions from reviews. In *HLT '05: Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL, Morristown, NJ, USA, pages 339–346.

Snyder, Benjamin and Regina Barzilay. 2007. Multiple aspect ranking using the good grief algorithm. In Candace L. Sidner, Tanja Schultz, Matthew Stone, and ChengXiang Zhai, editors, *HLT-NAACL*. The Association for Computational Linguistics, pages 300–307.

Titov, Ivan and Ryan McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In *Proc. of ACL-08: HLT*. ACL, Columbus, Ohio, pages 308–316.

Titov, Ivan and Ryan McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *WWW '08: Proc. of the 17th international conference on World Wide Web*. ACM, New York, NY, pages 111–120.

Turney, Peter. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. of 40th Annual Meeting of the Association for Computational Linguistics*. ACL, Philadelphia, Pennsylvania, USA, pages 417–424.

Zagibalov, Taras and John Carroll. 2008. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *COLING '08: Proc. of the 22nd International Conference on Computational Linguistics*. ACL, Morristown, NJ, USA, pages 1073–1080.

Zhu, X. and Z. Ghahramani. 2002. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02.