Using Aspect Extraction Approaches to Generate Review Summaries and User Profiles

Christopher Mitcheltree* **Veronica Wharton*** **Avneesh Saluja**
Airbnb AI Lab
San Francisco, CA, USA
`firstname.lastname@airbnb.com`

Abstract

Reviews of products or services on Internet marketplace websites contain a rich amount of information. Users often wish to survey reviews or review snippets from the perspective of a certain aspect, which has resulted in a large body of work on aspect identification and extraction from such corpora. In this work, we evaluate a newly-proposed neural model for aspect extraction on two practical tasks. The first is to extract canonical sentences of various aspects from reviews, and is judged by human evaluators against alternatives. A k-means baseline does remarkably well in this setting. The second experiment focuses on the suitability of the recovered aspect distributions to represent users by the reviews they have written. Through a set of review reranking experiments, we find that aspect-based profiles can largely capture notions of user preferences, by showing that divergent users generate markedly different review rankings.

1 Introduction

Aspect extraction has traditionally been associated with the sentiment analysis community (Liu, 2012; Pontiki et al., 2016), with the goal being to decompose a small document of text (e.g., a review) into multiple facets, each of which may possess their own sentiment marker. For example, a restaurant review may comment on the ambiance, service, and food, preventing the assignment of a uniform sentiment over the entire review. A common approach to aspect extraction is to treat the aspects as latent variables and utilize latent Dirichlet allocation (LDA; Blei et al. (2003)) to extract relevant aspects from a collection of documents in an unsupervised (Titov and McDonald, 2008; Brody and Elhadad, 2010) or semi-supervised (Mukherjee and Liu, 2012) fashion. Subsequent research has taken the latent variable approach further by encoding more complicated dependencies between aspects and sentiment (Zhao et al., 2010), or between aspects, ratings, and sentiment (Diao et al., 2014), using probabilistic graphical models (Koller and Friedman, 2009) to jointly learn the parameters.

However, it has been argued that the coherence of aspects extracted from the family of LDA-based approaches is low; words clustered together within a specific aspect are often unrelated, which can be attributed to the lack of word co-occurrence information in these models (Mimno et al., 2011), since conventional LDA assumes each word in a document is generated independently. Recently, He et al. (2017) proposed a neural attention-based aspect extraction (ABAE) approach, which like LDA, is an unsupervised model. The starting point is a set of word embeddings, where the vector representation of the word encapsulates co-occurrence[1]. The embeddings are used to represent a sentence as a bag-of-words, weighted with a self-attention mechanism (Lin et al., 2017), and learning amounts to encoding the resulting attention-based sentence embedding as a linear combination of aspect embeddings, optimized using an autoencoder formulation (§2). The attention mechanism thus learns to highlight words that will be pertinent for aspect identification.

In this work, we apply the ABAE model to a large corpus of reviews on Airbnb[2], an online marketplace for travel; users (guests) utilize the site to find accommodation (listings) all around the world, and a large number of these guests write reviews of the listing post-stay. We first provide additional details on the workings of the ABAE

*Equal contribution.

[1] words that co-occur with each other get mapped to points close to each other in the embedding space (Harris, 1968; Schütze, 1998).

[2] `www.airbnb.com`

Proceedings of NAACL-HLT 2018, pages 68–75
New Orleans, Louisiana, June 1 - 6, 2018. ©2017 Association for Computational Linguistics

model (§2). ABAE is then applied to two tasks: the first (§3.1) is to extract a representative sentence from a set of listing-specific reviews for a number of pre-defined aspects e.g., cleanliness and location, with the efficacy of extractive summarization evaluated by humans (§4.3). Surprisingly, we find that the k-means baseline performs very well on aspects that occur more frequently, but ABAE may be better for infrequent aspects.

In the second task (§3.2), we analyze the suitability of aspect embeddings to represent guest profiles. The hypothesis is that the content of guest reviews reveals the guest's preferences and priorities (Chen et al., 2015), and that these preferences correspond to extracted aspects. We investigate several ways to aggregate sentence-level aspect embeddings at the review and user levels and compute distances between user aspect and listing review embeddings, in order to personalize listing reviews by reranking them for each user. The correlation between guest profile distances (computed on pairs of guests) and review rank distances (computed on pairs of ordinal rankings over reviews) is then measured to evaluate our hypothesis (§4.4). We find a robust relationship between distances in the two spaces, with the correlation increasing at finer granularities like sentences compared to reviews or listings.

2 Background

To start, we provide a brief background of the ABAE model. For additional details, please refer to the original paper (He et al., 2017). At a high level, the ABAE model is an autoencoder that minimizes the reconstruction error between a weighted bag-of-words (BoW) representation of a sentence (where the weights are determined by a self-attention mechanism) and a linear combination of aspect embeddings. The linear combination represents the probabilities of the sentence belonging to each of the aspects.

The first step in ABAE is to compute the embedding $\mathbf{z}_s \in \mathbb{R}^d$ for a sentence s:

$$\mathbf{z}_s = \sum_{i=1}^{n} a_i \mathbf{e}_{w_i}$$

where $\mathbf{e}_{w_i}$ is the word embedding $\mathbf{e} \in \mathbb{R}^d$ for word w_i. As in the original paper, we use word vectors trained using the skip-gram model with negative sampling (Mikolov et al., 2013). The attention weights a_i are computed as a multiplicative self-attention model:

$$a_i = \text{softmax}(\mathbf{e}_{w_i}^{\mathrm{T}} \cdot \mathbf{M} \cdot \mathbf{y}_s)$$

$$\mathbf{y}_s = \sum_{i=1}^{n} \mathbf{e}_{w_i}$$

where $\mathbf{y}_s$ is simply the uniformly-weighted BoW embedding of the sentence, and $\mathbf{M} \in \mathbb{R}^{d\times d}$ is a learned attention model.

The next step is to compute the aspect-based sentence representation $\mathbf{r}_s \in \mathbb{R}^d$ in terms of an aspect embeddings matrix $\mathbf{T} \in \mathbb{R}^{K\times d}$, where K is the number of aspects:

$$\mathbf{p}_s = \text{softmax}(\mathbf{W} \cdot \mathbf{z}_s + \mathbf{b})$$

$$\mathbf{r}_s = \mathbf{T}^{\mathrm{T}} \cdot \mathbf{p}_s$$

where $\mathbf{p}_s \in \mathbb{R}^K$ is the weight (probability) vector over K aspect embeddings, and $\mathbf{W} \in \mathbb{R}^{K\times d}, \mathbf{b} \in \mathbb{R}^K$ are parameters of a multiclass logistic regression model.

The model is trained to minimize reconstruction error (using the cosine distance between $\mathbf{r}_s$ and $\mathbf{z}_s$) with a contrastive max-margin objective function (Weston et al., 2011). In addition, an orthogonality penalty term is added to the objective, which encourages the aspect embedding matrix $\mathbf{T}$ to produce diverse (orthogonal) aspect embeddings.

3 Tasks

To evaluate the utility of ABAE, we craft two methods of evaluation that mimic the practical ways in which aspect extraction can be used on a marketplace website with reviews.

3.1 Extractive Summarization

The first task is a direct evaluation of the quality of the recovered aspects: we use ABAE to select review sentences of a listing that are representative of a set of preselected aspects, namely cleanliness, communication, and location. "Cleanliness" refers to how clean the listing is, "communication" refers to communication between the listing host and the guest, and "location" refers to the qualities of or amenities in the listing's neighborhood. Refer to Table 3 for representative words for each aspect. Thus, aspect extraction is used to summarize listing reviews along several manually-defined topics.

We benchmark the ABAE model's extracted aspects against those from two baselines: LDA and k-means. For each experimental setup, the authors assigned one of four interpretable labels (corresponding to the identified aspects and the "other" category) to each unlabeled aspect by evaluating the 50 words most associated with that aspect[3]. LDA's topics are represented as distributions over words, so the most associated words correspond to those that occur with highest probability. For k-means and ABAE, each aspect is represented as a point in word embedding space[4], so we retrieve the 50 closest words to each point using cosine distance as a measure.

After aspect identification, we infer aspect distributions for the review sentences of an unseen test set of listings. For LDA, identification simply amounts to computing the (approximate) posterior over topic mixtures for a set of review sentences, and selecting the sentences with the highest probability in the specified aspect. For k-means, each sentence is represented as a uniformly-weighted BoW embedding, and we retrieve the sentences that are closest to the centroids that correspond to our preselected aspects. ABAE is similar, except the self-attention mechanism is applied to compute an attention-based BoW embedding, and we retrieve the sentences closest to the aspect embeddings corresponding to our aspects of interest. For some aspects (e.g., location and communication), there is a many-to-one mapping between the recovered word clusters and the aspect label. In these cases, we compute the average of the aspect embeddings, and the closest sentences to the resulting points are retrieved.

In our selection process, we retrieve the three most representative sentences (across all reviews of that listing) for each aspect. Three human annotators then evaluated the appropriateness of the selected aspect for each sentence via binary judgments. For example, an evaluator was presented with the sentence "Easy to get there from center of city by subway and bus.", along with the inferred aspect (*location*), for which a binary "yes/no" response suffices. Results aggregated by experimental setup and aspect are presented in §4.3.

3.2 Aspects as Profiles

An aspect extraction model provides a distribution over aspects for each sentence, and we can consider these distributions as interpretable sentence embeddings (since we can assign a meaning corresponding to an aspect for each of the extracted word clusters). These embeddings can be used to provide *guest profiles* by aggregating aspect distributions over review sentences that a user has written across different listings on the website. Such profiles arguably capture finer-grained information about guest preferences than an aggregate star rating across all aspects. Star ratings are also heavily positively-biased: more than 80% of reviews on Airbnb rate the maximum of 5 stars.

There are many conceivable ways to aggregate the sentence distributions, with some of the factors of variation being:

1. level of hierarchy: is the guest considered to be a bag-of-reviews (BoR), sentences (BoS), or words i.e., do we weight longer sentences or reviews with more sentences higher when computing the aggregated representation?
2. time decay: how do we treat more recently written reviews compared to earlier ones?
3. average or maximum: is the aggregate representation a (weighted) average, or should we consider the maximum value across each aspect and renormalize?

The same considerations also arise when computing the representations of the objects to be ranked e.g., a listing embedding as the aggregation of its component sentence embeddings.

For our evaluation (§4.4), guest profiles are computed by averaging distributions across a guest's review sentences uniformly (BoS), equivalent to a BoR representation weighted by the review length (number of sentences). We also experimented with a BoR representation using uniformly-weighted reviews, and the results are very similar to the BoS representation. We considered computing the guest profile by utilizing the maximum value (probability) of each aspect dimension across all review sentences written by the user and renormalizing the resulting embedding using the softmax function, but this approach resulted in high-entropic guest profiles with limited use downstream. More complex aggregation functions, like using an exponential moving average to upweight recent reviews, is an interesting future direction to explore.

4 Evaluation

We now look at the qualitative and quantitative performance of ABAE across the two tasks. After providing statistics on the review corpus that forms the basis of our evaluation, we qualitatively analyze the recovered aspects of the model,

[3] Inter-annotator agreements for each setup are provided in Table 3.

[4] The aspect embeddings in ABAE are initialized using the k-means centroids.

compared to k-means and LDA baselines. On a heldout evaluation set, human evaluators assessed whether the model-extracted aspects correspond to their understanding of the predefined ones by inspecting the top-ranked sentences for each aspect. Furthermore, the quality of the guest profile embeddings was evaluated by looking at the correlation between distances in the aspect space and the ordinal position of reviews on a given listing page, with the hypothesis that guests who write reviews with divergent content or aspects should receive rankings that are very different.

Our experiments were implemented using the pyTorch package[5]. Word vectors were trained using Gensim (Řehůřek and Sojka, 2010) with 5 negative samples, window size 5, and dimension 200, and Scikit-learn (Pedregosa et al., 2011) was used to run the k-means algorithm and LDA with the default settings. For ABAE, we used Adam with a learning rate of 0.001 (and the default β parameters) with a batch size of 50, 20 negative samples, and an orthogonality penalty weight of 0.1. All experiments were run on an Amazon AWS `p2.8xlarge` instance.

4.1 Datasets

The corpus was extracted from all reviews across all listings on Airbnb written between January 1, 2010 and January 1, 2017. We used spaCy[6] to segment reviews into sentences and remove non-English sentences. All sentences were subsequently preprocessed in the same manner as He et al. (2017), which entailed restricting the vocabulary to the 9,000 most frequent words in the corpus after stopword and punctuation removal. From the resulting set, we randomly sampled 10 million sentences across 5.8 million guests and 1.8 million listings to form a training set, and used the remaining unsampled sentences to select validation and test sets for the human evaluation (§4.3) and ranking (§4.4) experiments.

To select datasets for human evaluation, we identified all listings with at least 50 and at most 100 reviews in all languages and filtered out any *listing* in the training set, resulting in 900 listings which were split into validation and test sets. The validation set is used to select an appropriate number of aspects, by computing coherence scores (Mimno et al., 2011) as the number of aspects is varied in the ABAE model (§4.2). The test set was used to extract review sentences that were presented to our human evaluators; we ensured that every listing in the test set has at least 3 non-empty English review sentences.

For the ranking correlation experiments, we first identified users who had written at least 10 review sentences in our corpus and removed those users that featured in the training set from this list. We then selected 20 users uniformly at random to form our validation set i.e., to compute guest profiles for[7]. A subset of the human evaluation test set was used to compute the correlation between aspect space and ranking order distances; we selected all listings that had at least 20 review sentences, resulting in 69 listings for evaluation. Table 1 presents a summary of corpus statistics for all of the datasets used in this work.

Set	Task	Tokens	Sentences	Guests	Listings
Train	-	68.0mil	10.0mil	5.8mil	1.8mil
Val	§3.1	91,124	14,173	3719	721
Test	§3.1	21,069	3389	920	168
Val	§3.2	3189	543	20	202
Test	§3.2	13,925	2269	587	69

Table 1: Corpus statistics for the datasets that we use. All numbers are computed after preprocessing.

4.2 Recovered Aspects

Table 2 presents coherence scores for the ABAE model as we varied the number of aspects. Similar to He et al. (2017), we considered a "document" to be a sentence, but treating reviews as documents or all reviews of a listing as a document revealed similar trends. The table shows that coherence score improvements taper off after 30 aspects, so we chose this aspect value for further experiments.

Num.	**Num. Representative Words**			Sum
Aspects	10	30	50	
5	-125	-1106	-2829	-4060
10	-148	-1244	-3017	-4409
15	-126	-1069	-2656	-3851
30	-101	-760	-1917	-2778
40	-84	-701	-1765	-2550

Table 2: Coherence scores as a function of the number of aspects and the number of representative words used to compute the scores (higher is better). The summed values indicate significant improvement from 15 to 30 aspects. For details on computing coherence score, see Mimno et al. (2011).

Next, for each 30-aspect experimental setup, we identified the word clusters corresponding to the set of preselected aspects by labeling

[5] `http://pytorch.org/`

[6] `http://spacy.io/`

[7] The most prolific guest in this set had written 66 review sentences.

		Aspects		
Setup	Location	Cleanliness	Communication	Fleiss' κ
k-means	union, music, minute, dozen, quarter, chain, zoo, buffet, nord, theater (3 clusters, 10.2%)	master, conditioners, boiler, fabric, roll, smelling, dusty, shutter, dirty, installed (1 cluster, 3.6%)	welcomed, sorted, proactive, fix, checkin, prior, replied, process, communicator, ahead (3 clusters, 9.9%)	0.58
LDA	restaurant, location, flat, walk, away, back, short, minute, bus, come (4 clusters, 16.2%)	house, comfortable, clean, bed, beach, street, part, modern, apartment, cool (1 cluster, 3.5%)	helpful, arrival, wonderful, coffee, loved, use, warm, communication, friendly, got (4 clusters, 14.5%)	0.46
ABAE	statue, tavern, woodsy, street, takeaway, woodland, cathedral, specialty, idyllic, attraction (6 clusters, 18.4%)	clean, neat, pictured, immaculate, spotless, stylish, described, uncluttered, tidy, classy (1 cluster, 3.5%)	dear, u, responsive, greeted, instruction, communicative, sent, contract, attentive, key (3 clusters, 10.1%)	0.46

Table 3: Representative words for each aspect of interest across experimental setups, along with the number of clusters mapped to that aspect in parentheses as well as the percentage of validation set sentences assigned to that cluster (the remaining sentences were assigned to "Other"). For the aspects with multiple clusters, we select a roughly equal number of words from each cluster. Misspellings are deliberate.

each revealed cluster with a value from the set $\{cleanliness, communication, location, other\}$. Note that the mapping from clusters to identified aspects is many-to-one (i.e., multiple clusters for the same aspect were identified for two of the three aspects, namely location and communication.) In fact, the number of clusters associated with each aspect is a proxy for the frequency with which these aspects occur in the corpus. To verify this claim, we computed aspect-based representations (§3.1) for each sentence in the validation set used for comparing coherence scores, and utilized these representations to compute sentence similarities to each cluster, followed by a softmax in order to assign fractional counts i.e., a soft clustering approach. For each setup, Table 3 provides the top 10 words associated with each aspect, the number of clusters mapped to that aspect, and the number of validation sentences assigned to the aspect. The location and communication aspects are 3 to 6 times more prevalent than the cleanliness aspect.

Qualitatively, the ABAE aspects are more coherent, especially in the cleanliness aspect, and do not include irrelevant words (often verbs) that are not indicative of any conceivable aspect, like "got", "use", or "come". k-means selects relevant words to indicate the aspect, but the aspects are relatively incoherent compared to ABAE. LDA has a difficult time identifying relevant words, indicating the importance of the attention mechanism in ABAE. Interestingly, we found that the inter-annotator agreement (Fleiss' κ) was slightly higher for the k-means baseline, but all scores are in the range of moderate agreement.

		Aspects	
Setup	Loc	Clean	Comm
k-means	**0.85/0.68**	0.30/0.26	**0.62/0.43**
LDA	0.16/0.17	0.09/0.10	0.11/0.13
ABAE	0.45/0.46	**0.45/0.32**	0.41/0.35

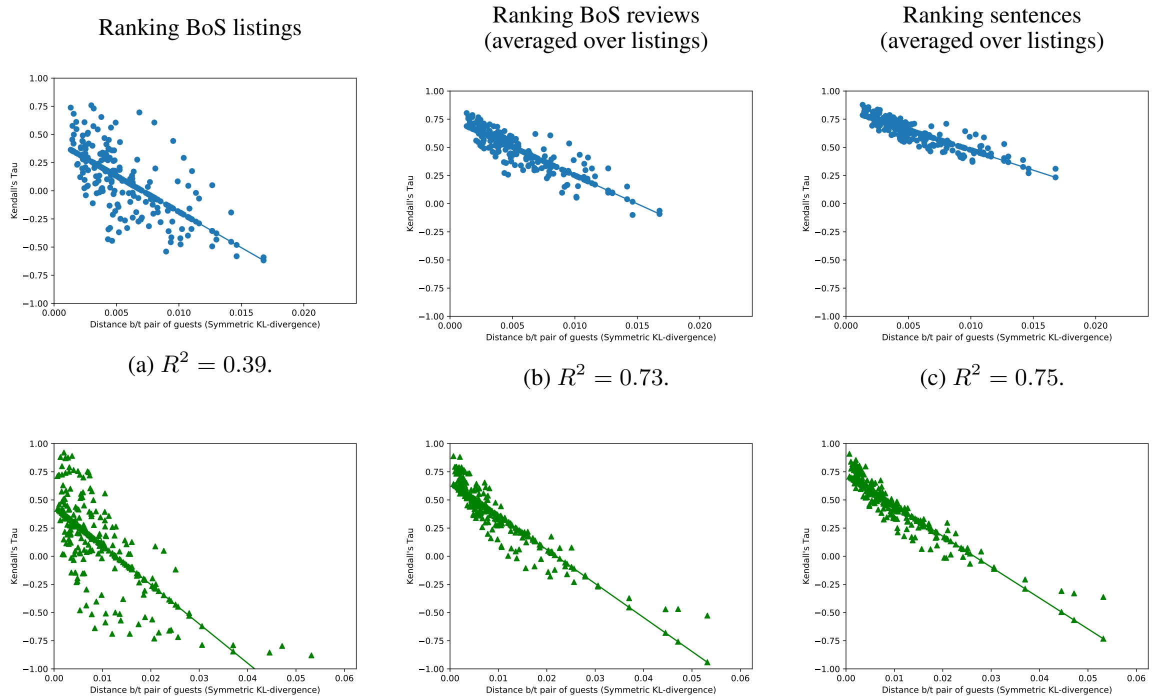
Table 4: Precision@1 and precision@3 for the extractive summarization task, as judged by our human evaluators.

4.3 Extracting Prototypical Sentences

Table 4 presents precision@1 and precision@3 results for each experimental setup-aspect pair, as evaluated by our human annotators. There are a total of 168 listings × 3 experimental setups × 3 aspects × 3 sentences per aspect $= 4536$ examples to evaluate; we set aside 795 examples to compute inter-annotator agreement, resulting in 2042 examples per annotator. Fleiss' $\kappa = 0.69$, which is quite high given the difficulty of the task[8].

The most surprising result is that the k-means baseline is actually the strongest performer in the location and communication aspects. Nonetheless, the result is encouraging since it suggests that, for some aspects of interest to us, a simple k-means approach and uniformly-weighted BoW embeddings suffices. It is interesting to note that the strong baseline performance occurs with the aspects that occur more frequently in the corpus, as discussed in §4.2, suggesting that ABAE is more useful with aspects that occur more rarely in our corpus (e.g., cleanliness). For future work, we propose to evaluate this hypothesis in more depth by applying the approaches in this paper to the long tail of rarer aspects. The disappointing performance of LDA shows that its lack of aware-

[8] The *communication* aspect (referring to host responsiveness and timeliness) is often easily confused with the friendliness of the host or staff.



(a) $R^2 = 0.39$.

(b) $R^2 = 0.73$.

(c) $R^2 = 0.75$.

(d) $R^2 = 0.46$.

(e) $R^2 = 0.81$.

(f) $R^2 = 0.84$.

Figure 1: Plots showing the relationship between distance in aspect space and ranking space, for guest profiles computed as bag-of-sentences representations. Figures 1a, 1b, and 1c are produced with the ABAE model, and figures 1d, 1e, while 1f are produced with the k-means model.

ness for word co-occurrence is damaging for aspect identification.

4.4 Review Ranking

Figure 1 presents results for the ranking correlation experiments with the ABAE and k-means models. The validation set is used to compute pairwise distances between all $\binom{20}{2} = 190$ guest pairs using the symmetric KL divergence, since guest profiles are probability distributions over aspects. This divergence forms the x-axis for our plots. We then rerank several objects of interest, and compute the rank correlation coefficient (Kendall's τ) between pairs of rankings; this coefficient forms the y-axis for our plots. Lastly, the correlation between the distance in aspect space (between pairs of user profiles) and the distance in ranking space (between pairs of rankings over objects, as measured by Kendall's τ) with R^2 values stated in the captions.

With the guest profiles, we ranked the following objects using the symmetric KL divergence:

1. listings, where each listing is represented as a BoS (similar results were achieved when considering each listing as a BoR).
2. reviews within a listing: for each guest pair and listing, we ranked the reviews using each guest's profile and computed Kendall's τ between the ranked pair of reviews. That score was then averaged over the 69 listings to yield a single score for each guest pair.
3. sentences within a listing: similar to reviews within a listing, except Kendall's τ was computed over ranked sentences. The averaging step was the same as above.

Since ABAE extracts aspects at the sentence-level, we would expect to see sentence-based representations result in higher correlations than other representations. Indeed, if we rank smaller units (i.e., sentences vs. listings), the correlation with distances in aspect space is higher (0.75 vs. 0.39 in the case of ABAE, 0.84 vs. 0.46 in the case of k-means). Interestingly, the correlation results are slightly better for k-means: the range of values for the pairwise distances (x-axis) is much larger, so it seems like the k-means guest profiles are better at capturing extremely divergent users, and the resulting ranking pairs are more divergent too. Table 5 presents an example of divergent rankings over review sentences for a given listing from two different guest profiles using the ABAE model.

Rank	Guest 1	Guest 2
1	Room is cozy and clean only the washroom feel a little bit old.	Within walking distance to Feng Chia Night Market yet quiet enough when it comes time to rest.
2	Clean and comfortable room for the lone traveller or couples.	Nice and clean place to stay, very near to Fengjia night market.
3	The room is very good, as good as on the photos, and also clean.	Overall my TaiChung trip was good and really convenient place to stay at Nami's place.
4	Nice and clean place to stay, very near to Fengjia night market.	Ia a great place to stay, clean.
5	Within walking distance to Feng Chia Night Market yet quiet enough when it comes time to rest.	Near feng jia night market.

Table 5: From the experiment in §4.4, ranked review sentences for two different guest profiles for the same listing using the ABAE model. The first guest's profile focuses on the listing interior and cleanliness aspects, whereas the second guest is more interested in location.

5 Conclusion

In this work, we evaluated a recently proposed neural-based aspect extraction model in several settings. First, we used the inferred sentence-level aspects to select prototypical review sentences of a listing for a given aspect, and evaluated this aspect identification/extractive summarization task using human evaluators benchmarked against two baselines. Interestingly, the k-means baseline does quite well on frequently-occurring aspects. Second, the sentence-level aspects were also used to compute user profiles by grouping reviews that individual users have written. We showed that these embeddings are effective in reranking sentences, reviews, or listings in order to personalize this content to individual users.

For future work, we wish to investigate alternative ways to aggregate and compute user profiles and compute distances between objects to rank and user profiles. We would also like to utilize human evaluators to judge the rankings produced in the review reranking experiments.

Acknowledgments

We thank the authors of the ABAE paper (He et al., 2017) for providing us with their code[9], which we used to benchmark our internal implementation. We are also grateful to our human evaluators.

References

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

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of NAACL-HLT*.

Li Chen, Guanliang Chen, and Feng Wang. 2015. Recommender systems based on user reviews: the state of the art. *User Modeling and User-Adapted Interaction*.

Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J. Smola, Jing Jiang, and Chong Wang. 2014. Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In *Proceedings of KDD*.

Zellig Harris. 1968. Mathematical structures of language. In *Interscience tracts in pure and applied mathematics*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of ACL*.

Daphne Koller and Nir Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of ICLR*.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR: abs/1310.4546*.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of EMNLP*.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proceedings of ACL*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning

[9] `https://github.com/ruidan/Unsupervised-Aspect-Extraction`

in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, Veronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Núria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of SemEval*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24:97–123.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of WWW*.

Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. Wsabie: Scaling up to large vocabulary image annotation. In *Proceedings of IJCAI*.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *Proceedings of EMNLP*.