Unsupervised Aspect Extraction

A report submitted for the course named Compiler Design(CS-320)

Kshitiz Kumar
Bachelor of Technology, V Semester
Roll : 16010103

Department of Computer Science and Engineering
Indian Institute of Information Technology , Senapati

Abstract

Aspect extraction is an important and challenging task in aspect-based sentiment analysis. Existing methods usually do not produce highly coherent aspects. In this project, I present a novel neural approach with the aim of discovering coherent aspects. The model improves coherence by exploiting the distribution of word co-occurrences through the use of neural word embeddings. Unlike topic models which typically assume independently generated words, word embedding models encourage words that appear in similar contexts to be located close to each other in the embedding space. In addition, one use an attention mechanism to de-emphasize irrelevant words during training, further improving the coherence of aspects.

Keywords : Aspect Extraction, word embedding , attention mechanism

Declaration

I declare that this submission represents my idea in my own words and where others' idea or words have been included, I have adequately cited and referenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/sources in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and can also evoke penal action from the sources which have thus not been properly cited or from proper permission has not been taken when needed.

(Signature)

(KSHITIZ KUMAR)

Date: (16010103)

Department of Computer Science & Engineering
Indian Institute of Information Technology Senapati

Mr. Himangshu Sharma Email: himangshu@iiitmanipur.ac.in

To Whom It May Concern

This is to certify that the report entitled **"Unsupervised Aspect Extraction"** submitted to by "Kshitiz Kumar", has been carried out under my supervision and that this work has not been submitted elsewhere for a degree,diploma or a course.

Signature of Supervisor

(Mr. Himangshu Sharma)

Acknowledgement

I am highly indebted to our faculty Mr Himangshu Sharma for his guidance and constant supervision as well as for providing necessary information regarding the project and also for their support in completing the project. I have taken efforts in this project. However, it would not have been possible without the help of all the faculties of Computer Science and Engineering Department,IIIT Manipur who took keen interest on my project work and guided me along, till the completion of my project work. I would like to extend my sincere thanks to all of them for encouraging us to do this project and made it a success.

- Kshitiz Kumar

Contents

List of Figures

1 Introduction

Aspect extraction is one of the key tasks in sentiment analysis. It aims to extract entity aspects on which opinions have been expressed. For example, in the sentence "The kebab was tender and melted in my mouth", the aspect term is "kebab". Two sub-tasks are performed in aspect extraction: (1) extracting all aspect terms (e.g., "kebab") from a review corpus, (2) clustering aspect terms with similar meaning into categories where each category represents a single aspect (e.g., cluster "kebab", "momo", "pasta", and "tomato" into one aspect food). Previous works for aspect extraction can be categorized into three approaches: rule-based, supervised, and unsupervised. Rule-based methods usually do not group extracted aspect terms into categories. Supervised learning requires data annotation and suffers from domain adaptation problems. Unsupervised methods are adopted to avoid reliance on labeled data needed for supervised learning.

In recent years, Latent Dirichlet Allocation (LDA) and its variants have become the dominant unsupervised approach for aspect extraction. LDA models the corpus as a mixture of topics (aspects), and topics as distributions over word types. While the mixture of aspects discovered by LDA-based models may describe a corpus fairly well, we find that the individual aspects inferred are of poor quality – aspects often consist of unrelated or loosely-related concepts. This may substantially reduce users' confidence in using such automated systems. There could be two primary reasons for the poor quality. Conventional LDA models do not directly encode word co-occurrence statistics which are the primary source of information to preserve topic coherence . They implicitly capture such patterns by modeling word generation from the document level, assuming that each word is generated independently. Furthermore, LDA-based models need to estimate a distribution of topics for each document. Review documents tend to be short, thus making the estimation of topic distributions more difficult.

In this work, I present a novel neural approach to tackle the weaknesses of LDA-based methods. I start with neural word embeddings that already map words that usually co-occur within the same context to nearby points in the embedding space . One then filter the word embeddings within a sentence using an attention mechanism and use the filtered words to construct aspect embeddings. The training process for aspect embeddings is analogous to autoencoders, where we use dimension reduction to extract the common factors among embedded sentences and reconstruct each sentence through a linear combination of aspect embeddings. The attention mechanism de-emphasizes words that are not part of any aspect, allowing the model to focus on aspect words.

In contrast to LDA-based models, the proposed method explicitly encodes word-occurrence statistics into word embeddings, uses dimension reduction to extract the most important aspects in the review corpus, and uses an attention mechanism to remove irrelevant words to further improve co- herence of the aspects.

2 Related Work

The problem of aspect extraction has been well studied in the past decade. Initially, methods were mainly based on manually defined rules. Some proposed to extract different product features through finding frequent nouns and noun phrases. They also extracted opinion terms by finding the synonyms and antonyms of opinion seed words through WordNet. Following this, a number of methods have been proposed based on frequent item mining and dependency information to extract product aspects . These models heavily depend on predefined rules which work well only when the aspect terms are restricted to a small group of nouns. Supervised learning approaches generally model aspect extraction as a standard sequence labeling problem. Rule-based models are usually not refined enough to categorize the extracted aspect terms. On the other hand, supervised learning requires large amounts of labeled data for training purposes.

Unsupervised approaches, especially topic models, have been proposed subsequently to avoid reliance on labeled data. Generally, the outputs of those models are word distributions or rankings for each aspect. Aspects are naturally obtained without separately performing extraction and categorization. Most existing works are based on variants and extensions of LDA .

Attention models have recently gained popularity in training neural networks and have been applied to various natural language processing tasks, including machine translation, sentence summarization, sentiment classification , and question answering . Rather than using all available information, attention mechanism aims to focus on the most pertinent information for a task. Unlike previous works, in this project, I want to apply attention to an unsupervised neural model.

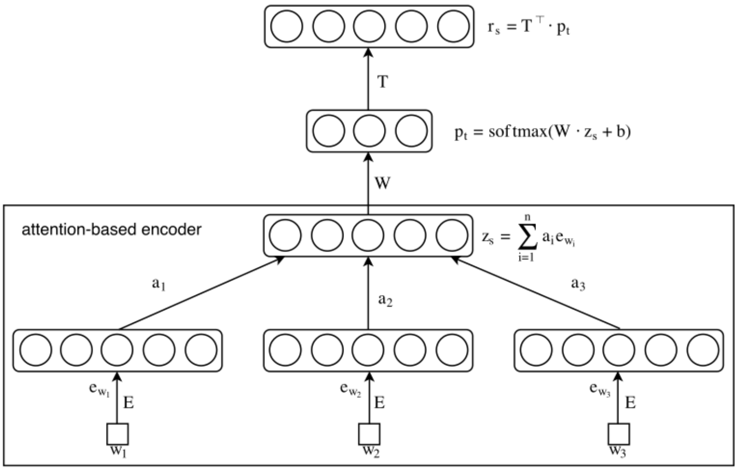


Figure 1: Structure of Unsupervised Aspect Extraction Model

3 Model Description

The ultimate goal is to learn a set of aspect embeddings, where each aspect can be interpreted by looking at the nearest words (representative words) in the embedding space. I begin by associating each word 'w' in the vocabulary with a feature vector e w ∈ R d . I used word embeddings for the feature vectors as word embeddings are designed to map words that often co-occur in a context to points that are close by in the embedding space . The feature vectors associated with the words correspond to the rows of a word embedding matrix E ∈ R V ×d , where V is the vocabulary size. I wanted to learn embeddings of aspects, where aspects share the same embedding space with words. This requires an aspect embedding matrix T ∈ R K×d , where K, the number of aspects defined, is much smaller than V . The aspect embeddings are used to approximate the aspect words in the vocabulary, where the aspect words are filtered through an attention mechanism.

Each input sample to ABAE is a list of indexes for words in a review sentence. Given such an input, two steps are performed as shown in Figure 1. First, I filter away non-aspect words by down-weighting them using an attention mechanism, and construct a sentence embedding from weighted word embeddings. Then, I try to reconstruct the sentence embedding as a linear com- bination of aspect embeddings. This process of dimension reduction and reconstruction, where model aims to transform sentence embeddings of the filtered sentences into their reconstructions with the least possible amount of distortion, preserves most of the information of the aspect words in the 'K' embedded aspects. I next describe the process in detail.

3.1 Sentence Embedding with Attention Mechanism

I construct a vector representation for each input sentence in the first step. In general, I want the vector representation to capture the most relevant information with regards to the aspect (topic) of the sentence. I define the sentence embedding as the weighted summation of word embeddings corresponding to the word indexes in the sentence.

For each word in the sentence, I compute a positive weight which can be interpreted as the probability that is the right word to focus on ,in order to capture the main topic of the sentence. The weight is computed by an attention model, which is conditioned on the embedding of the word as well as the global context of the sentence:

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

$$a_i = \frac{\exp(d_i)}{\sum_{j=1}^{n} \exp(d_j)}$$

$$d_i = \mathbf{e}_{w_i}^{\top} \cdot \mathbf{M} \cdot \mathbf{y}_s$$

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

where y_s is simply the average of the word embeddings, which I believe captures the global context of the sentence. $M \in R$ ($d \times d$) is a matrix mapping between the global context embedding and the word embedding and is learned as part of the training process. One can think of the attention mechanism as a two-step process. Given a sentence, I first construct its representation by averaging all the word representations. Then the weight of a word is assigned by considering two things. First, I filter the word through the transformation which is able to capture the relevance of the word to the 'K' aspects. Then I capture the relevance of the filtered word to the sentence by taking the inner product of the filtered word to the global context .

3.2 Sentence Reconstruction with Aspect Embeddings

I have obtained the sentence embedding. Now I describe how to compute the reconstruction of the sentence embedding. As shown in Figure 1, the reconstruction process consists of two steps of transitions, which is similar to an autoencoder. Intuitively, one can think of the reconstruction as a linear combination of aspect embeddings :

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

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

3.3 Training Attention-based Model

The model is trained to minimize the reconstruction error. For each input sentence, we randomly sample 'm' sentences from the training data as negative samples.Each negative sample is computed by averaging its word embeddings. The objective is to make the reconstructed embedding similar to the target sentence embedding while different from those negative samples.

3.4 Regularization Term

I hope to learn vector representations of the most representative aspects for a review dataset. However, the aspect embedding matrix may suffer from redundancy problems during training. To ensure the diversity of the resulting aspect embeddings, I added a regularization term to the objective function to encourage the uniqueness of each aspect embedding:

$$U(\theta) = \|\mathbf{T}_n \cdot \mathbf{T}_n^{\top} - \mathbf{I}\|$$

4 Experimental Setup

4.1 Dataset Description

We evaluate this method on two real-word datasets.

1. Citysearch corpus: This is a restaurant review corpus widely used by previous work, which contains over 50,000 restaurant reviews from Citysearch New York. These annotated sentences are used for evaluation of aspect identification. There are six manually defined aspect labels: Food, Staff, Ambience, Price, Anecdotes, and Miscellaneous.
2. BeerAdvocate: This is a beer review corpus, containing over 1.5 million reviews. A subset of 1,000 reviews, corresponding to 9,245 sen- tences, are annotated with five aspect labels: Feel, Look, Smell, Taste, and Overall.

Domain	#Reviews	#Labeled sentences
Restaurant	52,574	3,400
Beer	1,586,259	9,245

4.2 Previously Used Methods

To validate the performance of model ,I want to compare it against a number of baselines:

1. LocLDA : This method uses a standard implementation of LDA. In order to prevent the inference of global topics and direct the model towards rateable aspects, each sentence is treated as a separate document.
2. k-means:I initialized the aspect matrix by using the k-means centroids of the word embeddings.
3. SAS : This is a hybrid topic model that jointly discovers both aspects and aspect-specific opinions. This model has been shown to be competitive among topic models in discovering meaningful aspects.
4. BTM : This is a biterm topic model that is specially designed for short texts such as texts from social media and review sites. The major advantage of BTM over conventional LDA models is that it alleviates the problem of data sparsity in short documents by directly modeling the generation of unordered word-pair co-occurrences (biterms) over the corpus. It has been shown to perform better than conventional LDA models in discovering coherent topics.

5 Results

I want to describe the evaluation task and report the expected experimental results in this section. I want to evaluate the model on two criteria:

- Is it able to find meaningful and semantically coherent aspects?
- Is it able to improve aspect identification performance on real-world review datasets?

5.1 Aspect Quality Evaluation

Table below presents all 14 aspects for the restaurant domain. Compared to gold-standard labels, the inferred aspects are more fine- grained. For example, it can distinguish main dishes from desserts, and drinks from food.

Inferred Aspects	**Representative Words**	**Gold Aspects**
Main Dishes	beef, duck, pork, mahi, filet, veal	
Dessert	gelato, banana, caramel, cheesecake, pudding, vanilla	
Drink	bottle, selection, cocktail, beverage, pinot, sangria	Food
Ingredient	cucumber, scallion, smothered, stewed, chilli, cheddar	
General	cooking, homestyle, traditional, cuisine, authentic, freshness	
Physical Ambience	wall, lighting, ceiling, wood, lounge, floor	Ambience
Adjectives	intimate, comfy, spacious, modern, relaxing, chic	
Staff	waitstaff, server, staff, waitress, bartender, waiter	Staff
Service	unprofessional, response, condescending, aggressive, behavior, rudeness	
Price	charge, paid, bill, reservation, came, dollar	Price
Anecdotes	celebrate, anniversary, wife, fiance, recently, wedding	Anecdotes
Location	park, street, village, avenue, manhattan, brooklyn	
General	excellent, great, enjoyed, best, wonderful, fantastic	Misc.
Other	aged, reward, white, maison, mediocrity, principle	

5.1.1 Coherence Score

In order to objectively measure the quality of aspects, I use coherence score as a metric which has been shown to correlate well with human judgment .

A higher coherence score indicates a better aspect interpretability, i.e., more meaningful and semantically coherent. This indicates that neural word embedding is a better model for capturing co-occurrence than LDA, even for BTM which specifically models the generation of co-occurring word pairs.

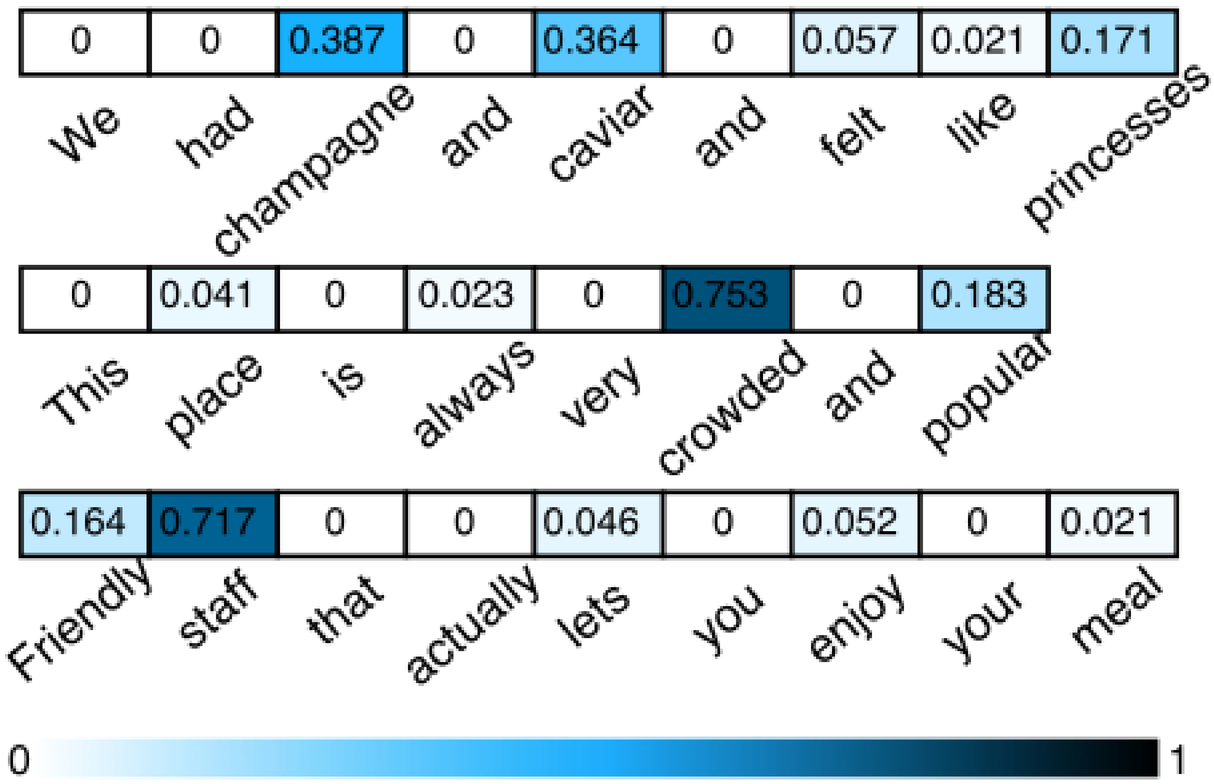
5.1.2 Real World Human Perception Example

As we want to discover a set of aspects that the human user finds agreeable, it is also necessary to carry out user evaluation directly. Following the experimental setting , I used a dummy data of recruitement of three human judges. Each aspect is labeled as coherent if the majority of judges assess that most of its top 50 terms coherently represent a product aspect. For a coherent aspect, each of its top terms is labeled as correct if and only if the majority of judges assess that it reflects the related aspect. One can observe that the user evaluation results correlate well with the coherence scores , where the model can substantially outperform all other models .

5.2 Aspect Identification

I evaluated the performance of sentence-level aspect identification on both domains using the annotated sentences . The evaluation criterion is to judge how well the predictions match the true labels. Given a review sentence, the model first assigns an inferred aspect label which corresponds to the highest weight . And one can then assign the gold-standard label to the sentence according to the mapping between inferred aspects and gold-standard labels.

For the restaurant domain, I followed the experimental settings to make the results comparable. To do that, (1) I only used the single-label sentences for evaluation to avoid ambiguity (about 83% of labeled sentences have a single label), and (2) I only evaluated on three major aspects, namely Food, Staff, and Ambience. The other aspects do not show clear patterns in either word usage or writing style, which makes these aspects very hard for even humans to identify.



5.3 Validating the Effectiveness of Attention Model

The weights learned by the model correspond very strongly with human intuition. In order to evaluate how attention model affects the overall performance of Aspect Extraction, one conduct experiments to compare Aspect Extraction with and without the attention layer and sentence embedding is calculated by averaging its word embeddings.

The results on the restaurant domain showed that the model achieves substantially higher precision and recall on all aspects which demonstrates the effectiveness of the attention mechanism.

6 Conclusion

I have presented Unsupervised Aspect Extraction (UAS) model, a simple yet effective neural attention model for aspect extraction. In contrast to LDA models, it explicitly captures word co-occurrence patterns and overcomes the problem of data sparsity present in review corpora. The experimental results demonstrated that the model not only learns substantially higher quality aspects, but also more effectively captures the aspects of reviews than previous methods.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In Proceedings of the 3rd International Conference on Learning Representations.

[2] David Blei, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3:993–1022.

[3] Samuel Brody and Noemie Elhadad. 2010. An unsu- pervised aspect-sentiment model for online reviews. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

[4] Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. 2014. Aspect extraction with automated prior knowledge learning. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

[5] Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.

[6] Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In 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.