Unsupervised Neural Aspect Extraction with Sememes

Ling Luo[1,5] , **Xiang Ao**[1,5*] , **Yan Song**[2] , **Jinyao Li**[3,5] , **Xiaopeng Yang**[4] ,
Qing He[1,5] and **Dong Yu**[2]

[1]Key Lab of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China
[2]Tencent AI Lab
[3]Institute of Software, Chinese Academy of Sciences
[4]David R. Cheriton School of Computer Science, Faculty of Mathematics, University of Waterloo
[5]University of Chinese Academy of Sciences
{luoling18s, aoxiang, heqing}@ict.ac.cn, {clksong, ljyyolia}@gmail.com, x335yang@uwaterloo.ca

Abstract

Aspect extraction relies on identifying aspects by discovering coherence among words, which is challenging when word meanings are diversified and processing on short texts. To enhance the performance on aspect extraction, leveraging lexical semantic resources is a possible solution to such challenge. In this paper, we present an unsupervised neural framework that leverages sememes to enhance lexical semantics. The overall framework is analogous to an autoenoder which reconstructs sentence representations and learns aspects by latent variables. Two models that form sentence representations are proposed by exploiting sememes via (1) a hierarchical attention; (2) a context-enhanced attention. Experiments on two real-world datasets demonstrate the validity and the effectiveness of our models, which significantly outperforms existing baselines.

1 Introduction

Aspect extraction is an essential component for aspect level sentiment analysis which is an important natural language processing task to identify peoples' opinions towards aspects of entities [Liu, 2012]. In general, there are two sub-tasks in aspect extraction, extracting aspect terms from an opinion corpus, e.g. identifying "*seaweed*" from "*The seaweed was chewy like gum*", and grouping subtle aspect terms into categories where each category denotes an individual aspect, e.g. cluster "*steak*" and "*seaweed*" into aspect "*food*".

Unsupervised approaches are one of the mainstream solutions. Among them, the common choices are latent Dirichlet allocation (LDA) based methods [Titov and McDonald, 2008; Brody and Elhadad, 2010; Zhao *et al.*, 2010; Mukherjee and Liu, 2012]. However, their performance is limited for two reasons. First, by exploiting word co-occurrences, it is hard to predict aspects from word topic (aspect) distributions because they are not highly differentiated. Second, the inaccurate topic distribution estimations from LDA-based models limit the performance on concise reviews.

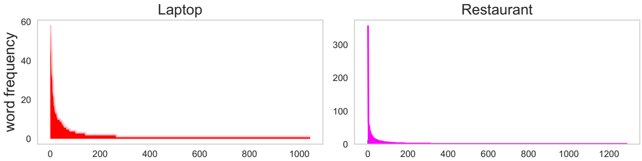

Figure 1: Statistics on aspect term frequency.

To this end, deep learning based models [Wang *et al.*, 2015; He *et al.*, 2017] were proposed. Yet they are still restrained from the following issues. First, a word in different contexts may have various senses, e.g., "roll" in two reviews "*This roll of Kodak film is of good quality*" and "*California rolls are awesome*" is different while it should be extracted as an aspect term in the latter review. Existing models do not explicitly distinguish different meanings of a word mainly because each word is equipped with a fixed embedding that is usually built upon word co-occurrences without considering its semantic information. Second, aspects might be represented in implicit manners. For example, "*my cellphone really fits my hand*" describes a positive sentiment on the "*size*" aspect, where the word "*fit*" is more indicative to "*size*". However, previous models are more likely to focus on words like "*cellphone*" because it is a domain-specific word while "*fit*" is more general. Third, most aspect terms are infrequent. Figure 1 exhibits the frequencies of the aspect terms on the dataset from SemEval Challenge 2014 task 4[1], and we observe an obvious long-tailed distribution on aspect terms' frequencies. However, existing neural models favor extracting frequent aspect terms instead of long-tailed ones.

Considering the manual process on aspect extraction, human being is skilled at solving these issues with exter-

[*] Corresponding author.

[1] This dataset, a benchmark for aspect-based sentiment classification task, includes the reviews from restaurant and laptop domain with 1, 288 and 1, 042 different annotated aspect terms, respectively.

same word in different meaning
implicit aspect
frequent aspect are detected better than rare aspect

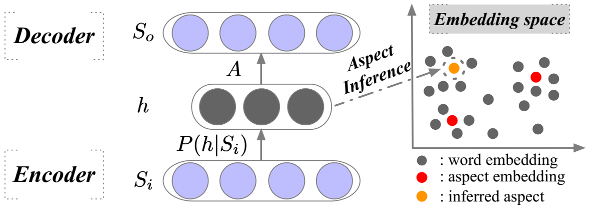

Figure 2: The Overall Framework.

nal semantic knowledge. To present such knowledge, sememes [Bloomfield, 1926] are minimum semantic units of word meanings and usually annotated in lexical semantic resources. It could be effective to find the real aspect terms especially infrequent ones by utilizing such semantic resources to obtain latent semantic information and get the meaning of the context behind the obscure and various expressions.

In doing so, in this paper, we leverage sememes from external lexical semantic resources and introduce an unsupervised neural framework to incorporate neural models and unlabeled data effectively. The overall framework is analogous to an autoencoder, which takes a sentence representation as input. The sentence representation is then reconstructed by a linear combination of *aspect embeddings* and a latent variable sampling from a learnt distribution. Based on the framework, we propose two models which leverage sememes to form the input sentence representations in different ways. The first model, namely, **A**spect **E**xtraction with **S**ememe **A**ttentions (AE-SA), has a hierarchical sememe attention layer that obtains a sentence representation by emphasizing correlated word senses on the lexical level. The second model, namely, **A**spect **E**xtraction via **C**ontext-enhanced **S**ememe **A**ttentions (AE-CSA) adopts an RNN to perform global encoding of sentences and concatenates with the sememe attention layer to effectively explore related word senses. Experiments on two large public review datasets have verified the effectiveness of our models. and also show that sememes can greatly help discover infrequent aspects, where AE-SA and AE-CSA outperform the previous state-of-the-art models. The contributions can be summarized as follows,

- Our models utilize lexical semantics to discover latent semantic information behind implicit and various expressions for aspect extraction.
- We propose a sememe attention structure to represent word meanings and such structure is proved to be useful in aspect extraction, especially for extracting infrequent aspects.
- We add an RNN structure to the sememe attention, which learns the sequential information of the contexts and help the explorations of real aspect terms.

2 The Proposed Model

In this section, we describe our models incorporating sememes for unsupervised aspect extraction. First, we detail the overall framework. Then, we present our two models AE-SA and AE-CSA, which consist of different network structures to get sentence representations.

2.1 The Overall Framework

The overall framework illustrated in Figure 2 consists of an encoding and a decoding process. In the encoding process, the input sentence representation S_i is firstly reduced to a compressed latent variable $h \in \mathbb{R}^K$, where K is the number of pre-defined aspects[2]. During decoding, the output sentence representation S_o is constructed by a linear combination of the latent variable h and an aspect embedding matrix $A \in \mathbb{R}^{K \times d}$, where d is the embedding size. The aspect embedding is aligned with the word embeddings so as to be inferred by performing the nearest neighbor search in the embedding space.

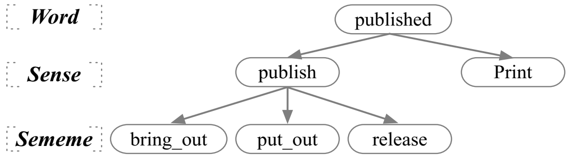

Figure 3: The structure of sememes.

Encoder. S_i is encoded to a latent variable $h \in \mathbb{R}^K$ and h can be seen as a probability vector over K aspects. We sample h from a continuous latent distribution, which is assumed as a Gaussian $\mathcal{N}(h|\mu, \sigma^2)$ in our models where $\mu = l_1(S_i)$ and $\log\sigma = l_2(S_i)$. l_1, l_2 can be any type of neural network and we simply use Multi-Layer Perceptrons (MLP) in our implementation. $h \sim \mathcal{N}(\mu, \sigma^2)$ is sampled as the output of the encoder (c.f. Eq. 1) where ϵ is a random value from a normal distribution.

$$h = \mu + \sigma \odot \epsilon \tag{1}$$

Decoder. A reconstructed sentence representation S_o is derived by a linear combination of the aspect embedding matrix $A \in \mathbb{R}^{K \times d}$ with the latent variable h (c.f. Eq. 2) in the decoding process. A is learned during the training process with each row representing an aspect embedding. As shown in Figure 2, the semantic meaning of each aspect can be inferred by its nearest words from the embedding space since the aspect embedding is aligned with the word embeddings.

$$S_o = A^T \cdot h, \tag{2}$$

Training and test. The loss function L of our models consists of a training objective function J and a regularization term U. J is designed to minimize the reconstruction error of S_i and S_o and differ S_o from q randomly selected negative samples $\{N_1, \ldots, N_q\}$. We compute the sum of fixed word embeddings as the j-th negative sample representation N_j. We utilize contrastive max-margin for J [Rush *et al.*, 2015], meanwhile U is aimed to enhance the diversity and avoid redundancy of aspect embeddings. In detail,

$$L = J + \lambda U, \tag{3}$$

$$J = \sum_{m \in D} \sum_{j=1}^{q} \max\left(0, 1 - S_o^m S_i^m + S_o^m N_j^m\right), \tag{4}$$

$$U = \parallel A_n \cdot A_n^T - I \parallel, \tag{5}$$

where D represents the training dataset and λ is a hyperparameter. A_n represents the normalization on each row of

[2] Here the aspect indicates topical aspect category, which is distinct from aspect term. For example, "seaweed" is an aspect term in "The seaweed was chewy like gum", and "Food" could be its topical aspect category. We will continue to use the notion of "aspect" as topical aspect category hereafter unless other specified.

A, while I is the identity matrix. During the test, a sentence S_i' is encoded to h', then we perform topical aspect inference by analyzing the weights in h' since it indicates the probability over the pre-defined topical aspects.

2.2 Aspect Extraction with Sememe Attentions (AE-SA)

Sememes [Bloomfield, 1926] annotated on lexical semantic resources are minimum semantic units of word meanings. In detail, a *word* may have multiple *senses* and a *senses* consists of several *sememes*. See Figure 3 for an example. The word "*published*" has more than one senses ("*publish*","*print*", etc.), and the sense "*publish*" is composed of some sememes ("*bring_out*","*put_out*","*release*").

AE-SA adopts a hierarchical sememe attention layer to form a better word and sentence representations by de-emphasizing irrelevant sememes and focusing on more correlative ones. Figure 4 details its sememe attention layers.

Given a sentence $s = \{w_1, \ldots, w_m\}$, we compute the average of word embeddings as E_{avg} and construct an initial sentence representation S by,

$$S = \sum_{i=1}^{m} \texttt{softmax}(\tanh(e_i^T \cdot E_{avg}))e_i, \tag{6}$$

where $e_i \in \mathbb{R}^d$ is the word embedding of w_i. The following is a hierarchical sememe attention mechanism driven by S. Through this, we derive a new sentence representation S_i by,

$$\widetilde{x}_{t,i} = \sum_{j=1}^{l_i} \texttt{softmax}(\tanh({x_{i,j}^t}^T \cdot S))x_{i,j}^t, \tag{7}$$

$$e_t' = \sum_{i=1}^{n} \texttt{softmax}(\tanh(\widetilde{x}_{t,i}^T \cdot S))\widetilde{x}_{t,i}, \tag{8}$$

$$S_i = \sum_{t=1}^{m} \texttt{softmax}(\tanh({e_t'}^T \cdot M \cdot S))e_t'. \tag{9}$$

The representation of the i-th sense in the t-th word is a weighted sum of its sememes $\{\mathbf{x}_{i,1}^t, \ldots, \mathbf{x}_{i,l_i}^t\}$ (c.f. Eq. 7). Then, the new embedding of the t-th word in sentence s is computed by aggregating different senses $\{\widetilde{\mathbf{x}}_{t,1}, \ldots, \widetilde{\mathbf{x}}_{t,n}\}$ with an attention mechanism (c.f. Eq. 8). Such hierarchical sememe attention attempts to expand senses of each word and capture its meaning within specific contexts. Next, we construct a new sentence representation S_i by strengthening relevant words with another attention mechanism (c.f. Eq. 9) where $M \in \mathbb{R}^{d \times d}$ is a trainable transformation matrix and d is the dimension of word and sentence vector.

2.3 Aspect Extraction via Context-enhanced Sememe Attentions (AE-CSA)

So far, with the sememe attention layer, AE-SA tends to obtain lexical semantic information and explores possible meanings of each word, which can extract real aspects behind the obscure and various expressions. Nevertheless, exploration on lexical level without considering the overall meaning of certain contexts might bring noises. To help the sememe attention explore reasonable lexical semantic information of words, we add an RNN-based structure to AE-SA and get the upgraded model, AE-CSA. Compared with AE-SA, AE-CSA utilize lexical semantic information to construct the input sentence with the help of words' sequential relations.

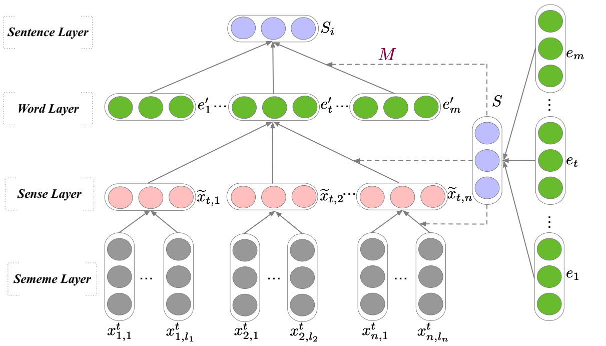

Figure 4: The structure of sememe attention.

In particular, apart from getting a sentence representation S_i via the sememe attention layer (c.f. Eq. 9), the original sentence $s = \{w_1, \ldots, w_m\}$ is also fed into an RNN to generate the hidden representation h_{rnn}. Then h_{rnn} and S_i jointly generate a new sentence S_i' via an MLP (c.f. Eq. 10). S_i' can be regarded as a kind of distorted sentence representation represented via the sememe attention, which is guided by the sequential and semantic constraints of the sentence. The new input sentence is then reconstructed in AE-CSA.

$$S_i' = \tanh(W^T \cdot (S_i \oplus h_{rnn}) + b) \tag{10}$$

where $W \in \mathbb{R}^{(d+d') \times d}$ and b are parameters to be learned. $\oplus$ stands for vector concatenation, $h_{rnn} \in \mathbb{R}^{d'}, S_i', S_i \in \mathbb{R}^d$, where d' and d are dimension of the hidden vector of RNN and the sentence vector, respectively. Intuitively, h_{rnn} may suggest the sememe attention layer which words are more important and should be determined as the appropriate word senses by influencing the training of the transformation matrix M (c.f. Eq. 9). S_i' is the derived sentence representation which is fed to the auto-encoder as Section 2.1.

3 Experiment Setup

3.1 Datasets

We conduct extensive experiments on two real-world datasets to evaluate our model. The statistics are shown in Table 1.

Citysearch corpus. It contains over $50,000$ restaurant reviews from Citysearch New York. An annotated subset with $3,400$ sentences are used for evaluation [Ganu *et al.*, 2009] with six manually defined aspect labels: *Food, Staff, Ambience, Price, Anecdotes* and *Miscellaneous*. We name this corpus **Restaurant** dataset hereafter.

BeerAdvocate. This is a beer review corpus provided in [McAuley *et al.*, 2012] containing more than 1.5 million reviews. A subset of $1,000$ reviews consisting of $9,245$ sentences, are annotated as five aspect labels *Feel, Look, Smell, Taste*, and *Others*. We name this dataset **Beer** hereafter.

3.2 Baselines

The baselines include representative unsupervised models. **SAS** [Mukherjee and Liu, 2012] and **BTM** [Yan *et al.*, 2013]

concatentae 200 vector with 500 vector and use dense to get vector of size 200

Dataset	Reviews #	Labeled Sentences #
Restaurant	52,574	3,400
Beer	1,586,259	9,245

Table 1: Statistics of the two datasets.

are variants of LDA-based models. Deep neural models are **SERBM** [Wang *et al.*, 2015] and **ABAE** [He *et al.*, 2017]. For a fair comparison, we extend ABAE and derive **ABAE-SEM** by averaging sememes embeddings to initialize the corresponding word embeddings. We also degrade our model by removing the attentions on sememes for ablation test. We denote it as **AE-S**, where the sentence embedding is an average of word embeddings, every word embedding is the mean of its sense embeddings and each sense is the mean of corresponding sememes.

3.3 Implementation

In the preprocessing, punctuations, stop words and words appearing less than 10 times in the corpus are removed. The NLTK pos-tagger[3] is used to annotate part-of-speech information for each word within specific contexts. We initialize the word embedding matrix by word2vec [Mikolov *et al.*, 2013] trained on the experimental datasets, and the embedding size is set to 200. The word vocabulary size is set to $9,000$ for Restaurant and $11,000$ for Beer, which can cover 96.64% and 98.84% contents on these two datasets, respectively. We utilize WordNet [Miller, 1995] to obtain word sememes. Specifically, each word is aggregated by 2 senses consisting of 5 sememes. Each sememe is represented by a set of words in their lemma forms, and its embedding is computed by averaging the corresponding word vectors in our experiments. The words in the two vocabularies can be fully covered by WordNet, hence every word can be expanded with ten sememes. 66.81% and 68.07% words representing sememes can be found in our vocabularies, respectively, and the rest OOV sememe words are padded with the original word itself. Following [He *et al.*, 2017], we set the number of aspects for both datasets to 14 and utilize the centroids of k-means clusters to initialize aspect embedding matrix A. Word embeddings are fixed during training. Adam [Kingma and Ba, 2014] is employed as the optimizer with learning rate of 0.001. Orthogonality penalty weight λ is set to 2 on Restaurant and 2.5 on Beer, respectively. For both datasets, the number of negative samples q is 20, the dimensions of S, S_i, S_i' are 200 and the hidden size of RNN structure h_{rnn} is 500 . LSTM is adopted as the RNN structure for AE-CSA and our models use the same set of parameters.

4 Experimental Results

As aspect embeddings are aligned with the word embedding space, we can infer the label of each aspect from its neighbor words in the embedding space and then manually build a mapping from the inferred aspects to the gold-standard labels following [Zhao *et al.*, 2010; Brody and Elhadad, 2010; Wang *et al.*, 2015; He *et al.*, 2017]. Remember that 14 distinct aspect embeddings are learnt on each dataset. Hence we infer these 14 aspects on the two datasets and map them to 6 and 5 gold-standard labels, respectively. For example, if an aspect embedding's nearest surrounding words in the embedding space are "*espresso*", "*martini*", "*sangria*", we can infer it as "*Drinks*" and then generalize it to "*Food*" in the Restaurant dataset. The results of SAS and SERBM are taken from [Wang *et al.*, 2015], and the results of BTM and ABAE are taken from [He *et al.*, 2017].

Aspect	Model	Precision	Recall	F1
Food	SAS	0.867	0.772	0.817
	BTM	0.933	0.745	0.816
	SERBM	0.891	0.854	0.872
	ABAE	**0.953**	0.741	0.828
	ABAE-SEM	0.870	0.914	0.892
	AE-S (Ours)	0.846	0.914	0.879
	AE-SA (Ours)	0.883	0.923	0.902
	AE-CSA (Ours)	0.903	**0.926**	**0.914**
Staff	SAS	0.774	0.556	0.647
	BTM	0.828	0.579	0.677
	SERBM	0.819	0.582	0.680
	ABAE	0.802	0.728	0.757
	ABAE-SEM	0.793	0.730	0.760
	AE-S (Ours)	0.757	0.653	0.701
	AE-SA (Ours)	**0.844**	0.705	0.768
	AE-CSA (Ours)	0.804	**0.756**	**0.779**
Ambience	SAS	0.780	0.542	0.640
	BTM	0.813	0.599	0.685
	SERBM	0.805	0.592	0.682
	ABAE	**0.815**	0.698	0.740
	ABAE-SEM	0.769	0.713	0.740
	AE-S (Ours)	0.759	0.677	0.716
	AE-SA (Ours)	0.771	0.715	0.742
	AE-CSA (Ours)	0.768	**0.773**	**0.770**

Table 2: Aspect inference results on the Restaurant dataset.

via evaluation

4.1 Quantitative Analysis on Aspect Inference

We first conduct on a quantitative analysis to evaluate the performance on sentence-level aspect inference. With the annotated sentences in both datasets, we can compute how well the predictions match the true labels. To perform the prediction, we first assign an inferred aspect label according to the highest weight in h. Then, we generalize this inferred aspect to a gold-standard label according to a manually created mapping.

Table 2 exhibits the results on Restaurant. Following [He *et al.*, 2017], we only evaluate on aspects $\{$*Food, Staff, Ambience*$\}$ out of $\{$*Food, Staff, Ambience, Price, Anecdotes, Miscellaneous*$\}$ and use sentences with single label to avoid ambiguity. AE-SA as well as AE-CSA significantly outperform the compared models on the Recall and F_1 measures. It proves the effectiveness of the proposed models structured with sememes on sentence-level aspect inference.

Table 3 displays the results on Beer dataset. In addition to the 5 gold-standard aspect labels $\{$*Feel, Taste, Smell, Look, Others*$\}$, we follow [He *et al.*, 2017] to form a combined aspect "*Taste+Smell*" for a fair comparison. In Table 3, we observe that AE-CSA outperforms the previous models on 5 out of the 6 aspects on F_1 scores. Though AE-CSA cannot surpass the compared methods on every aspect, we still find the following interesting observations. As "*Taste*" and "*Smell*" are very similar and many words can be used interchangeably to describe both aspects, e.g., "bitter", "spicy" "sweet", etc. can describe either "*Taste*" or "*Smell*" of beer. So previous studies perform not very well on "*Taste*" and "*Smell*", and no one can exceed 0.6 on F_1 scores. While for our AE-CSA, since it perceives the fine-grained semantics with the context-

[3] https://www.nltk.org/_modules/nltk/tag.html

Aspect	Model	Precision	Recall	F1
Feel	SAS	0.783	0.695	0.730
	BTM	0.892	0.687	0.772
	ABAE	0.806	0.824	0.816
	ABAE-SEM	0.886	**0.830**	0.818
	AE-S (Ours)	0.813	0.772	0.792
	AE-SA (Ours)	0.904	0.753	0.821
	AE-CSA (Ours)	0.909	0.757	**0.826**
Taste	SAS	0.543	0.496	0.505
	BTM	0.616	0.467	0.527
	ABAE	**0.637**	0.358	0.456
	ABAE-SEM	0.596	0.621	0.608
	AE-S (Ours)	0.589	0.577	0.583
	AE-SA (Ours)	0.589	0.629	0.609
	AE-CSA (Ours)	0.566	**0.741**	**0.641**
Smell	SAS	0.336	0.673	0.404
	BTM	0.541	0.549	0.527
	ABAE	0.483	0.744	**0.575**
	ABAE-SEM	0.434	**0.750**	0.550
	AE-S (Ours)	0.617	0.482	0.499
	AE-SA (Ours)	0.599	0.518	0.555
	AE-CSA (Ours)	**0.608**	0.541	0.572
Taste+Smell	SAS	0.804	0.759	0.769
	BTM	0.885	0.760	0.815
	ABAE	**0.897**	0.853	0.866
	ABAE-SEM	0.882	0.854	0.868
	AE-S (Ours)	0.839	0.830	0.834
	AE-SA (Ours)	0.887	0.860	0.874
	AE-CSA (Ours)	0.889	**0.875**	**0.882**
Look	SAS	0.958	0.705	0.806
	BTM	0.953	0.854	0.872
	ABAE	**0.969**	0.882	0.905
	ABAE-SEM	0.930	0.925	0.928
	AE-S (Ours)	0.903	0.912	0.908
	AE-SA (Ours)	0.938	**0.922**	0.930
	AE-CSA (Ours)	0.945	0.918	**0.932**
Others	SAS	0.618	0.664	0.619
	BTM	**0.699**	0.715	0.700
	ABAE	0.654	0.828	0.725
	ABAE-SEM	0.620	0.771	0.687
	AE-S (Ours)	0.572	0.833	0.678
	AE-SA (Ours)	0.623	0.836	0.714
	AE-CSA (Ours)	0.647	**0.838**	**0.730**

Table 3: Aspect inference results on the Beer dataset.

enhanced sememe attentions, it achieves 0.641 on F_1 on the "*Taste*" aspect. By looking closer to the top representative words from the results of AE-CSA on "*Taste*" and "*Smell*", we observe "*smell, aroma*" in "*Smell*" and "*mouthfeel, well-carbonated*" in "*Taste*", which are clearly different.

Moreover, ABAE-SEM performs slightly better than original ABAE with the help of sememes. Without the attention layers, AE-S exhibits imperfects compared with our proposed models. It runs even worse than ABAE on some aspects. These observations could also give insights on the effectiveness of the proposed sememe attentions.

4.2 Qualitative Analysis on Aspect Embeddings

As AE-CSA generally achieves the best performance on aspect inference among all the compared methods in the quantitative analysis, next we conduct a qualitative analysis to look closer to its aspect embeddings. In particular, we demonstrate all the 14 inferred aspects from the results of AE-CSA on Restaurant with their representative words, and the corresponding gold-standard aspect labels in Table 4. We observe the inferred aspects are fine-grained and their representative words are coherent and coupled. For example, it can discover the aspect terms like "*brooklyn*", "*houston*" and "*manhattan*" which are "*Location*" names and distinguish them from the words describing conceptual "*Place*" such as "*balcony*", "*station*" and "*entrance*". Meanwhile, concrete and infrequent representative words can be retrieved, e.g. "*risotto*"[4]. We conjecture it is because the context-enhanced sememe attentions that can discover fine-grained and latent information which contribute to better exploring aspect terms.

Inf. Aspects	Rep. Words	Gd. Aspects
Ingredient	gravy, cilantro, sesame	Food
Main Dishes	antipasti, seafood, risotto	
Drinks	espresso, champagne, martini	
Mental Feeling	agree, overrated, disappointing	
Staff	doorman, waitress, cashier	Staff
Service	courteous, overbearing, attentive	
Adjectives	terrific, incredible, fabulous	Ambience
Physical	lighting, banquette, ceiling	
Price	cheaper, pricy, charging	Price
Anecdotes	thanksgiving, valentine, easter	Anecdotes
Actions	plan, return, suggest	Misc.
Location	brooklyn, houston, manhattan	
Place	balcony, station, entrance	
Others	artisinal, congrats, marcus	

Table 4: Example aspects inferred by AE-CSA on the Restaurant dataset, with their representative words and the corresponding gold-standard aspect labels. The left column shows 14 manually labeled inferred aspects. The middle column shows the representative words selected from top 5 nearest words of each aspect embedding. The right column presents the gold-standard aspects.

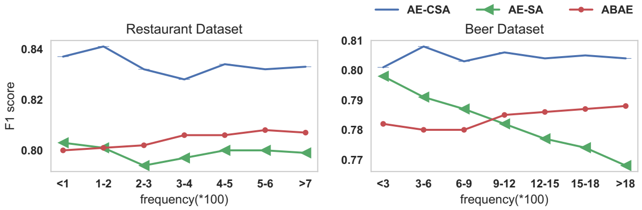

Figure 5: The F1 score of different frequency intervals of aspect terms on both datasets.

Next we design another experiment to testify the effectiveness of sememe attentions w.r.t different term frequencies. Specifically, we extract the highest attended word as the aspect term from each sentence in the test set. Then we get the inferred aspect for that sentence by querying the nearest aspect with the extracted word in the embedding space, and map them to the gold-standard labels. Finally, these aspect terms are divided into several intervals by their frequencies.

We measure the average F_1 score of 3 aspects on Restaurant and 6 aspects on Beer for each interval, and investigate the performances of some compared models. Figure 5 displays the results. First, AE-CSA performs better than the compared models on either frequent or infrequent aspect terms on both datasets. Meanwhile, the performance gap between AE-CSA and ABAE is getting smaller with aspect term frequency increasing. It exactly reflects the benefits of lexical semantic resources because frequent words usually have enough samples to learn effective representations while the low-frequent ones need external supports. As frequency increases, the effect of this support becomes insubstantial. Second, AS-SA performs worse than ABAE as frequency grows. It might be that though the exploitation of sememes attention is helpful for discovering the infrequent aspect terms, the performance may decrease without the constraint of the se-

[4] Rice cooked usually in meat or seafood stock and seasoned.

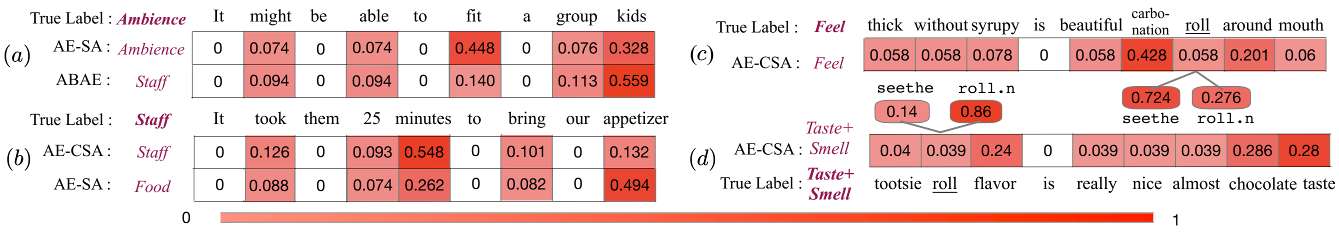

Figure 6: Visualization of weight allocations.

quential encoding of the original sentence.

4.3 Case Study

Finally we visualize the weight allocations of AE-SA vs ABAE, AE-CSA vs AE-SA and attentions in sense layer.

AE-SA vs ABAE. Given sentence shown as Fig. 6 (a), ABAE mainly focuses on the word "*kids*" and incorrectly infers the aspect as "*Staff*". While AE-SA correctly assigns "*Ambience*" to it with the highly attended word "*fit*". Though "*fit*" seems not to be an explicit aspect term, one group of its sememes "*suit, accommodate*" in the WordNet may facilitate its correlation to the correct aspect.

AE-CSA vs AE-SA. Given sentence shown as Fig. 6 (b), it shows that AE-CSA focuses more on "*minutes*" while AE-SA attends to extract "*appetizer*". The sentence generally describes the service is slow, so the ground truth is supposed to be "*Service*" which is mapped to the gold-standard aspect label as "*Staff*". The major reason for the difference is that the transformation matrix M of these two approaches are different. Note that M is trainable in our model as shown in Eq. 9. We argue it is the RNN-based structure of AE-CSA that improves the aspect inference task by influencing the training of M to assign more weights to the appropriate words.

Attention in sense layer. Sememes can extend semantics but bring noise as well. The balance between semantic injection and word sense disambiguation can be underpinned by the RNN structure in AE-CSA that captures the overall sentence meaning and helps the attention layers to explore exact sememes. As shown in Figure 6 (c) and (d), the word "*roll*" has various attended senses in sentence. Specifically, word "*roll*" mostly consists of the "*seethe*" meaning in "*carbonation roll around mouth*", while it represents a kind of food as a noun in sentence "*tootsie roll flavor is really nice*". We argue it is sememes constructed with RNN encoder that discover fine-grained information based on right meaning of the context and contribute to better exploring aspect terms.

5 Related Work

Aspect extraction. Aspect Extraction is important for sentiment analysis [Wang *et al.*, 2017a; Luo *et al.*, 2018], and existing work can be categorized into three types: rule-based, supervised, and unsupervised approaches. Rule-based methods [Somasundaran and Wiebe, 2009; Qiu *et al.*, 2011] exploited frequent pattern to extract product features. Supervised methods [Jin *et al.*, 2009; Li *et al.*, 2010; Choi and Cardie, 2010; Wang *et al.*, 2016; Wang and Pan, 2018] were typically built based on sequence labeling methods such as hidden Markov models (HMM) and conditional random fields (CRF), which required large amounts of labeled data for training. LDA-based methods [Titov and McDonald, 2008; Brody and Elhadad, 2010; Zhao *et al.*, 2010; Mukherjee and Liu, 2012; Chen *et al.*, 2014] were representative methods of unsupervised solutions. [Wang *et al.*, 2015] proposed Restricted Boltzmann Machines based model to extract aspects and relevant sentiments simultaneously. [Yin *et al.*, 2016] used dependency path embeddings and [Wang *et al.*, 2017b] proposed multi-layer attentions while exploiting indirect relations between terms. [Li and Lam, 2017] utilized a memory interaction structure [He *et al.*, 2017] adopted attention mechanism with word embeddings to improve aspect coherence.

Sememe representation learning. Recently, sememes from lexical semantic resources are utilized to recognize different word senses in various contexts and applied in many NLP tasks. [Niu *et al.*, 2017], [Song *et al.*, 2017] and [Zeng *et al.*, 2018] encoded it to improve Chinese word representations and lexicon expansion. [Li *et al.*, 2018; Qi *et al.*, 2018; Jin *et al.*, 2018] study the lexical sememe prediction task with external resources. In this paper, we devise different sememe structures and investigate their effectiveness for aspect extraction. To the best of our knowledge, we are the first to apply sememes in unsupervised aspect extraction tasks.

6 Conclusion

In this paper, we proposed unsupervised neural models incorporating sememes from lexical semantic resources for aspect extraction, where sememes are utilized by a hierarchical attention mechanism and a context-enhanced attention mechanism, respectively. Experiments showed that all models utilizing sememes improve aspect extraction in contrast to existing models because sememes help explore latent semantic information behind implicit and various expressions of sentences. Particularly, the context-enhanced sememe attention model performs better on identifying real aspects within the specific contexts and raising on the long-tail aspect terms.

Acknowledgements

This work is partially supported by the National Key Research and Development Program of China under Grant No. 2017YFB1002104, the National Natural Science Foundation of China under Grant No. 91846113, U1811461, 61602438, 61573335, CCF-Tencent RhinoBird Young Faculty Open Research Fund No. RAGR20180111. This work is also funded in part by Ant Financial through the Ant Financial Science Funds for Security Research. Xiang Ao is also supported by Youth Innovation Promotion Association CAS.

References

[Bloomfield, 1926] Leonard Bloomfield. A set of postulates for the science of language. *Language*, 1926.

[Brody and Elhadad, 2010] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *ACL*, 2010.

[Chen *et al.*, 2014] Zhiyuan Chen, Arjun Mukherjee, and Bing Liu. Aspect extraction with automated prior knowledge learning. In *ACL*, 2014.

[Choi and Cardie, 2010] Yejin Choi and Claire Cardie. Hierarchical sequential learning for extracting opinions and their attributes. In *ACL*, 2010.

[Ganu *et al.*, 2009] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. Beyond the stars: Improving rating predictions using review text content. In *WebDB*, 2009.

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

[Jin *et al.*, 2009] Wei Jin, Hung Hay Ho, and Rohini K Srihari. A novel lexicalized hmm-based learning framework for web opinion mining. In *ICML*, 2009.

[Jin *et al.*, 2018] Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. Incorporating chinese characters of words for lexical sememe prediction. In *ACL*, 2018.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Computer Science*, 2014.

[Li and Lam, 2017] Xin Li and Wai Lam. Deep multi-task learning for aspect term extraction with memory interaction. In *EMNLP*, 2017.

[Li *et al.*, 2010] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. Structure-aware review mining and summarization. In *COLING*, 2010.

[Li *et al.*, 2018] Wei Li, Xuancheng Ren, Damai Dai, Yunfang Wu, Houfeng Wang, and Xu Sun. Sememe prediction: Learning semantic knowledge from unstructured textual wiki descriptions. *arXiv preprint arXiv:1808.05437*, 2018.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[Luo *et al.*, 2018] Ling Luo, Xiang Ao, Feiyang Pan, Jin Wang, Tong Zhao, Ningzi Yu, and Qing He. Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention. In *IJCAI*, 2018.

[McAuley *et al.*, 2012] Julian McAuley, Jure Leskovec, and Dan Jurafsky. Learning attitudes and attributes from multi-aspect reviews. In *ICDM*, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 1995.

[Mukherjee and Liu, 2012] Arjun Mukherjee and Bing Liu. Aspect extraction through semi-supervised modeling. In *ACL*, 2012.

[Niu *et al.*, 2017] Yilin Niu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Improved word representation learning with sememes. In *ACL*, 2017.

[Qi *et al.*, 2018] Fanchao Qi, Yankai Lin, Maosong Sun, Hao Zhu, Ruobing Xie, and Zhiyuan Liu. Cross-lingual lexical sememe prediction. In *EMNLP*, 2018.

[Qiu *et al.*, 2011] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.

[Rush *et al.*, 2015] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for sentence summarization. In *EMNLP*, 2015.

[Somasundaran and Wiebe, 2009] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. In *ACL*, 2009.

[Song *et al.*, 2017] Yan Song, Chia-Jung Lee, and Fei Xia. Learning word representations with regularization from prior knowledge. In *CoNLL*, 2017.

[Titov and McDonald, 2008] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *WWW*, 2008.

[Wang and Pan, 2018] Wenya Wang and Sinno Jialin Pan. Transition-based adversarial network for cross-lingual aspect extraction. In *IJCAI*, 2018.

[Wang *et al.*, 2015] Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. Sentiment-aspect extraction based on restricted boltzmann machines. In *ACL*, 2015.

[Wang *et al.*, 2016] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Recursive neural conditional random fields for aspect-based sentiment analysis. In *EMNLP*, 2016.

[Wang *et al.*, 2017a] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, 2017.

[Wang *et al.*, 2017b] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *AAAI*, pages 3316–3322, 2017.

[Yan *et al.*, 2013] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *WWW*, 2013.

[Yin *et al.*, 2016] Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. Unsupervised word and dependency path embeddings for aspect term extraction. In *IJCAI*, 2016.

[Zeng *et al.*, 2018] Xiangkai Zeng, Cheng Yang, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. Chinese liwc lexicon expansion via hierarchical classification of word embeddings with sememe attention. 2018.

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