A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis

Qingnan Jiang[1], Lei Chen[1], Ruifeng Xu[2,3], Xiang Ao[4], Min Yang[1]*
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]Department of Computer Science, Harbin Institute of Technology (Shenzhen)
[3]Peng Cheng Laboratory
[4]Institute of Computing Technology, Chinese Academy of Sciences
jqnthomask@gmail.com, lei.chen@siat.ac.cn, xuruifeng@hit.edu.cn
aoxiang@ict.ac.cn, min.yang@siat.ac.cn

Abstract

Aspect-based sentiment analysis (ABSA) has attracted increasing attention recently due to its broad applications. In existing ABSA datasets, most sentences contain only one aspect or multiple aspects with the same sentiment polarity, which makes ABSA task degenerate to sentence-level sentiment analysis. In this paper, we present a new large-scale Multi-Aspect Multi-Sentiment (MAMS) dataset, in which each sentence contains at least two different aspects with different sentiment polarities. The release of this dataset would push forward the research in this field. In addition, we propose simple yet effective CapsNet and CapsNet-BERT models which combine the strengths of recent NLP advances. Experiments on our new dataset show that the proposed model significantly outperforms the state-of-the-art baseline methods[1].

1 Introduction

Aspect-based sentiment analysis (ABSA) aims at identifying the sentiment polarity towards the specific aspect in a sentence. An target aspect refers to a word or a phrase describing an aspect of an entity. For example, in the sentence "*The decor is not special at all but their amazing food makes up for it*", there are two aspect terms "*decor*" and "*food*", and they are associated with *negative* and *positive* sentiment respectively.

Recently, neural network methods have dominated the study of ABSA since these methods can be trained end-to-end and automatically learn important features. (Wang et al., 2016) proposed to learn an embedding vector for each aspect, and these aspect embeddings were used to calculate the attention weights to capture important information with regard to the given aspects. (Tang et al., 2016b) developed the deep memory network to compute the importance degree and text representation of each context word with multiple attention layers. (Ma et al., 2017) introduced the interactive attention networks (IAN) to interactively learn attentions in contexts and targets, and generated the representations for target and context words separately. (Xue and Li, 2018) proposed to extract sentiment features with convolutional neural networks and selectively output aspect related features for classification with gating mechanisms. Subsequently, Transformer (Vaswani et al., 2017) and BERT based methods (Devlin et al., 2018) have shown high potentials on ABSA task. There are also several studies attempting to simulate the process of human reading cognition to further improve the performance of ABSA (Lei et al., 2019; Yang et al., 2019).

So far, several ABSA datasets have been constructed, including SemEval-2014 Restaurant Review dataset, Laptop Review dataset (Pontiki et al., 2014) and Twitter dataset (Dong et al., 2014). Although these three datasets have since become the benchmark datasets for the ABSA task, most sentences in these datasets consist of only one aspect or multiple aspects with the same sentiment polarity (see Table 1)[2], which makes aspect-based sentiment analysis degenerate to sentence-level sentiment analysis. Based on our empirical observation, the sentence-level sentiment classifiers without considering aspects can still achieve competitive results with many recent ABSA methods (see TextCNN and LSTM in Table 3). On the other hand, even advanced ABSA methods trained on these datasets can hardly distinguish the sentiment polarities towards different aspects in the sentences that contain multiple aspects and multiple sentiments.

*Min Yang is corresponding author

[1]Data and code can be found as: https://github.com/siat-nlp/MAMS-for-ABSA

[2]ATSA and ACSA represent aspect-term and aspect-category sentiment analysis, respectively.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 6280–6285,
Hong Kong, China, November 3–7, 2019.

Dataset	Type	Categ.	Size	MM_size
Restaurant	ATSA	4	4827	1283
Restaurant	ACSA	4	4738	454
Laptop	ATSA	4	3012	604
Twitter	ATSA	3	6940	6

Table 1: Statistics of existing datasets for ABSA. Size and MM_size represent the total number of instances and multi-apsect multi-sentiment instances in the dataset. Each multi-apsect multi-sentiment instance contains multiple aspects with different sentiment polarities.

With the goal of advancing and facilitating research in the field of aspect-based sentiment analysis, in this paper, we present a new Multi-Aspect Multi-Sentiment (MAMS) dataset. In MAMS dataset, each sentence consists of at least two aspects with different sentiment polarities, making the proposed dataset more challenging compared with existing ABSA datasets. Considering merely the sentence-level sentiment of the sentence will fail to achieve good performance on MAMS dataset. We empirically evaluate the state-of-the-art ABSA methods on MAMS dataset, the poor results demonstrate that the proposed MAMS dataset is more challenging than the SemEval-2014 Restaurant Review dataset.

We analyze the properties of recent ABSA methods, and propose new capsule networks (denoted as CapsNet and CapsNet-BERT) to model the complicated relationship between aspects and contexts, which combine the strengths of recent NLP advances. Experimental results show that the proposed methods achieve significantly better results than the state-of-the-art baseline methods on MAMS and SemEval-14 Restaurant datasets.

Our main contributions are summarized as follows: (1) We manually annotate a large-scale multi-aspect multi-sentiment dataset, preventing ABSA degenerating to sentence-level sentiment analysis. The release of it would push forward the research of ABSA. (2) We propose a novel capsule network based model to learn the complicated relationship between aspects and contexts. (3) Experimental results show that the proposed method achieves significantly better results than the state-of-the-art baseline methods.

2 Dataset Construction

Data Collection Similar to SemEval-2014 Restaurant Review dataset (Pontiki et al., 2014), we annotate sentences from the Citysearch New York dataset collected by (Ganu et al., 2009). We split each document in the corpus into a few sentences, and remove the sentences consisting more than 70 words.

Dataset		Pos.	Neg.	Neu.	Total.
ATSA	Train	3380	2764	5042	11186
	Train (small)	1089	892	1627	3608
	Validation	403	325	604	1332
	Test	400	329	607	1336
ACSA	Train.	1929	2084	3077	7090
	Train (small)	1000	1100	1613	3713
	Validation	241	259	388	888
	Test	245	263	393	901

Table 2: Statistics of MAMS dataset.

Data Annotation We create two versions of MAMS dataset for two subtasks of aspect-based sentiment analysis: aspect-term sentiment analysis (ATSA) and aspect-category sentiment analysis (ACSA).

For ATSA, we invited three experienced researchers who work on natural language processing (NLP) to extract aspect terms in the sentences and label the sentiment polarities with respect to the aspect terms. The sentences that consist of only one aspect term or multiple aspects with the same sentiment polarities are deleted. We also provide the start and end positions in a sentence for each aspect term.

For ACSA, we pre-defined eight coarse aspect categories: *food*, *service*, *staff*, *price*, *ambience*, *menu*, *place* and *miscellaneous*. Five aspect categories are adopted in SemEval-2014 Restaurant Review Dataset. We add three more aspect categories to deal with some confusing situations. Three experienced NLP researchers were asked to identify the aspect categories described in given sentences and determine the sentiment polarities towards these aspect categories. We only keep the sentences which consist of at least two unique aspect categories with different sentiment polarities.

Dataset Analysis The statistics of MAMS dataset for ATSA and ACSA are reported in Table 2. MAMS consists of 13,854 instances for ATSA and 8,879 instances for ACSA, which is 2.87 and 1.87 times of SemEval-2014 Restaurant Review dataset respectively. In MAMS, the sentences contain 2.62 aspect terms and 2.25 aspect categories in average. All sentences in MAMS contains multiple aspects with different sentiment polarities. On the contrary, the ra-

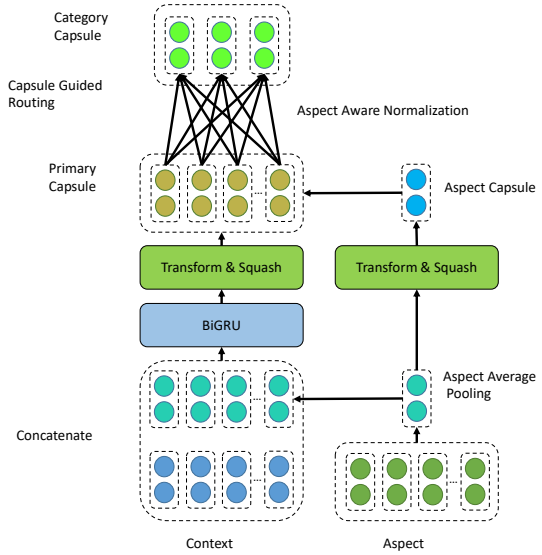

Figure 1: Model overview.

tio of multi-aspect multi-sentiment instances in Restaurant-ATSA, Restaurant-ACSA, Laptop and Twitter datasets are 26.58%, 9.58%, 20.05%, and 0.09% respectively, making the existing datasets less challenging since non-MAMS instances could be classified correctly without effectively modeling the relationships between contexts and aspects. For fair comparison with SemEval-14 Restaurant dataset, we also provide a small version of MAMS dataset (called MAMS-small) by sampling training instances from the MAMS training set, which has the same number of training instances with the SemEval-14 Restaurant dataset.

3 Methodology

We use D to denote the collection of sentences in the training data. Given a sentence $S = \{w_1^s, ..., w_n^s\}$, an aspect term $A^t = \{w_1^a, ..., w_m^a\}$ or an aspect category A^c, aspect-level sentiment classification aims to predict the sentiment polarity $y \in \{1, ..., C\}$ of sentence S with respect to A^t or A^c. Here, w denotes a specific word, n and m are the lengths of the sentence and aspect term, C represents the number of sentiment categories.

As illustrated in Figure 1, the proposed model consists of an embedding layer, an encoding layer, a primary capsule layer and a category capsule layer.

3.1 Embedding Layer

In the embedding layer, we convert the sentence S into word embeddings $\mathbf{E}$. For ACSA task, aspect category embedding $\mathbf{a}$ are randomly initialized and learned during training, while for ATSA task, aspect embedding $\mathbf{a}$ are computed as average pooling over aspect word embeddings. We get the aspect-aware sentence embedding $\mathbf{E}^{sa}$ by concatenating the aspect embedding $\mathbf{a}$ with each word embedding in S: $\mathbf{E}_i^{sa} = [\mathbf{E}_i; \mathbf{a}]$

3.2 Encoding Layer

We convert aspect-aware sentence representation $\mathbf{E}^{sa}$ into contextualized representation $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_n]$ through bidirectional gated recurrent unit (BiGRU) (Cho et al., 2014) and residual connection (He et al., 2016).

$$\mathbf{H} = \text{BiGRU}(\mathbf{E}^{sa}) + \mathbf{E}^{sa} \quad (1)$$

3.3 Primary Capsule Layer

In the primary capsule layer, we get primary capsules $\mathbf{P} = [\mathbf{p}_i, ...\mathbf{p}_n]$ and aspect capsule $\mathbf{c}$ through linear transformation and squashing activation.

$$\mathbf{p}_i = \text{squash}(\mathbf{W}^p \mathbf{h}_i + \mathbf{b}^p) \quad (2)$$

$$\mathbf{c} = \text{squash}(\mathbf{W}^a \mathbf{a} + \mathbf{b}^a) \quad (3)$$

where $\mathbf{W}^p$, $\mathbf{b}^p$, $\mathbf{W}^a$ and $\mathbf{b}^a$ are learnable parameters. The squash function is defined as:

$$\text{squash}(\mathbf{s}) = \frac{\|\mathbf{s}\|^2}{1 + \|\mathbf{s}\|^2} \frac{\mathbf{s}}{\|\mathbf{s}\|} \quad (4)$$

Aspect Aware Normalization Due to the variable lengths of sentences, the number of primary capsules sent to upper layer capsules varies from sentence to sentence, leading to unstable training procedure. Extremely long sentences make the squash activation saturate and result in high confidence for all categories; while very short sentences in contrast will lead to low confidence for all categories. To alleviate this problem, we propose the aspect aware normalization that utilizes aspect capsule to select important primary capsules, and normalize primary capsule weights $\mathbf{u}$ by:

$$u_i = \frac{\exp(\mathbf{p}_i \mathbf{W}^n \mathbf{c})}{\sum_{j=1}^{n} \exp(\mathbf{p}_j \mathbf{W}^n \mathbf{c})} \quad (5)$$

where $\mathbf{W}^n$ is a learnable parameter.

Capsule Guided Routing Original dynamic routing mechanism (Sabour et al., 2017) suffer from inefficient training due to the iteration procedure of routing. And there is no upper layer information used to guide the routing process, which

makes dynamic routing work like a self-directed process.

Instead of computing coupling coefficients between primary capsules and category capsules during routing process, we design a capsule-guided routing mechanism, which leverages prior knowledge about the sentiment categories to guide the routing process effectively and efficiently. Specifically, we use a set of sentiment capsules to store prior knowledge about the sentiment categories. Let $\mathbf{G} \in R^{C \times d}$ be the sentiment matrix, which is initialized with the averaged embeddings of sentiment words. C is the number of sentiment categories, and d is the dimension of the sentiment embedding. We can get the sentiment capsules $\mathbf{Z} = [\mathbf{z}_1, ..., \mathbf{z}_C]$ by applying squash activation over sentiment matrix and compute the routing weights $\mathbf{w}$ by calculating the similarity between primary capsules and sentiment capsules:

$$\mathbf{z}_i = \text{squash}(\mathbf{G}_i) \tag{6}$$

$$w_{ij} = \frac{\exp(\mathbf{p}_i \mathbf{W}^r \mathbf{z}_j)}{\sum_{k=1}^{n} \exp(\mathbf{p}_i \mathbf{W}^r \mathbf{z}_k)} \tag{7}$$

3.4 Category Capsule Layer

Based on the normalization weights and routing weights, the final category capsules $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_C]$ can be calculated as:

$$\mathbf{v}_j = \text{squash}(s \sum_{i=1}^{n} w_{ij} u_i \mathbf{p}_i) \tag{8}$$

where s is a learnable scale parameter to scale the connection weights to a suitable level.

Following (Sabour et al., 2017), we use the margin loss as loss function for aspect-based sentiment classification:

$$L = \sum_{k=1}^{C} T_k \max(0, m^+ - \|\mathbf{v}_k\|)^2 + \lambda(1 - T_k) \max(0, \|\mathbf{v}_k\| - m^-)^2 \tag{9}$$

where $T_k = 1$ if and only if a category k is present. m^+ and m^- are the margin hyper-parameters. λ control the loss for absent categories. In our experiments, m^+, m^- and λ is set to 0.9, 0.1, 0.6, respectively.

3.5 CapsNet-BERT

To utilize the features learned from large-scale corpus, we design the CapsNet-BERT model which combines the strength of BERT and capsule networks. We replace the embedding layer and encoding layer of CapsNet with pre-trained BERT. The CapsNet-BERT model takes "[CLS] sentence [SEP] aspect [SEP]" as input, which computes the deep representations of sentences and aspects with pre-trained BERT. We then feed the sentence and aspect representations into capsule layers and predict the corresponding sentiment polarities.

4 Experiments

4.1 Experimental Setup

Experimental Data In order to evaluate the effectiveness of our model, we conduct experiments on the two MAMS datasets and SemEval-14 Restaurant Review (Pontiki et al., 2014) dataset. All models share the same data pre-processing procedure, and use the same pre-trained word embeddings.

Baseline Methods We compare CapsNet with various baselines. (1) LSTM based models: TD_LSTM (Tang et al., 2016a), AT_LSTM (Wang et al., 2016), ATAE_LSTM (Wang et al., 2016), BiLSTM enhanced with attention mechanism, IAN (Ma et al., 2017) and AOA_LSTM (Huang et al., 2018); (2) CNN based model: GCAE (Xue and Li, 2018); (3) Attention based models: MemNet (Tang et al., 2016b) and AEN (Song et al., 2019). To verify the effectiveness of combining capsule networks and BERT, we also compare the performance of CapsNet-BERT with vanilla BERT model. Vanilla BERT model adopts "[CLS] sentence [SEP] aspect [SEP]" as input and predicts the sentiment polarity of the sentence towards the given aspect.

Implementation Details In all experiments, we use 300-dimentional word vectors pre-trained by GloVe (Pennington et al., 2014) to initialize the word embedding vectors for non-BERT models. The capsule size is set to 300. The batch sizes are set to 64 and 32 for CapsNet and CapsNet-BERT respectively. We use Adam optimizer (Kingma and Ba, 2015) to train our models. The learning rates are set to 0.0003 and 0.00003 for CapsNet and CapsNet-BERT respectively. We run all models for 5 times and report the average results on the test datasets. We fine-tune the hyper-parameters for all baselines on the validation set.

Method	ATSA			ACSA		
	MAMS	MAMS-small	Restaurant	MAMS	MAMS-small	Restaurant
TextCNN	52.694	51.736	75.928	48.856	48.814	81.418
LSTM	52.486	50.134	75.552	48.546	48.346	82.076
TD_LSTM	74.596	73.582	75.630	-	-	-
AT_LSTM	77.558	73.028	76.200	66.436	65.128	83.100
ATAE_LSTM	77.054	71.708	77.200	70.634	66.814	**84.000**
BiLSTM+Attn	76.166	70.720	77.356	66.304	66.262	80.514
IAN	76.602	71.168	78.600	-	-	-
AOA_LSTM	77.260	72.290	**81.200**	-	-	-
AEN	66.722	61.706	80.980	-	-	-
MemNet	64.568	62.694	80.950	63.288	62.818	77.862
GCAE	77.588	73.246	77.280	72.098	66.968	79.350
CapsNet	**79.776**	**73.860**	80.786	**73.986**	**67.128**	83.554
BERT	82.218	79.440	84.460	78.292	75.316	90.442
CapsNet-BERT	**83.391**	**80.910**	**85.934**	**79.461**	**76.366**	**91.375**
CapsNet-DR	79.434	71.398	77.768	69.036	65.638	81.904
CapsNet-BERT-DR	82.970	80.092	84.646	79.132	75.634	90.736

Table 3: Experimental results on two MAMS datasets and SemEval-2014 Restaurant Review dataset for both ATSA and ACSA subtasks.

4.2 Experimental Results and Analysis

Experimental results are reported in Table 3. From Table 3 we draw the following conclusions. First, sentence-level sentiment classifiers (TextCNN and LSTM) achieve competitive results on SemEval-14 Restaurant Review dataset but perform poorly on MAMS datasets. This verifies that MAMS datasets can alleviate the task degeneration problem encountered in Restaurant dataset for ABSA. Second, most advanced and complex ABAS methods, which achieve impressive results on Restaurant dataset, perform poorly on the MAMS-small dataset. This verifies that MAMS (small) is more challenging than SemEval-14 Restaurant Review dataset. Third, attention based models without effectively modeling word order (e.g., MemNet and AEN) perform worst on MAMS since they lose word order information and cannot identify which part of context describes the given aspect. Fourth, CapsNet outperforms non-BERT baselines on 4 of 6 datasets, showing the potential of applying capsule networks to aspect-based sentiment analysis task. In addition, CapsNet-BERT performs significantly better than other models including BERT, indicating that combining capsule network and BERT can obtain additional improvement compared to vanilla BERT.

4.3 Ablation Study

To analyze the effectiveness of the proposed capsule-guided routing mechanism, we conduct ablation study that replace capsule-guided routing by dynamic routing in both CapsNet and CapsNet-BERT, resulting in CapsNet-DR and CapsNet-BERT-DR. From Table 3 (last two rows) we can see that capsule-guided routing boosts the performance of CapsNet and CapsNet-BERT on all the datasets.

5 Conclusion

In this paper, we present MAMS, a challenge dataset for aspect-based sentiment analysis, in which each sentence contains multiple aspects with different sentiment polarities. The proposed MAMS dataset could prevent aspect-level sentiment classification degenerating to sentence-level sentiment classification, which might push forward the researches on aspect-based sentiment analysis. In addition, we propose a simple yet effective capsule networks that significantly outperforms compared methods.

Acknowledgements

Min Yang is partially supported by SIAT Innovation Program for Excellent Young Researchers (No. Y8G027) and sponsored by CCF-Tencent Open Research Fund. Xiang Ao is partially supported by the National Natural Science Foundation of China (No. 61602438), CCF-Tencent RhinoBird Young Faculty Open Research Fund (No. RAGR20180111), and Youth Innovation Promotion Association CAS.

References

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers)*, volume 2, pages 49–54.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. *international conference on learning representations*.

Zeyang Lei, Yujiu Yang, Min Yang, Wei Zhao, Jun Guo, and Yi Liu. 2019. A human-like semantic cognition network for aspect-level sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6650–6657.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *IJCAI'17 Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. *neural information processing systems*, pages 3856–3866.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. *international conference on computational linguistics*, pages 3298–3307.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *neural information processing systems*, pages 5998–6008.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. *meeting of the association for computational linguistics*, 1:2514–2523.

Min Yang, Qingnan Jiang, Ying Shen, Qingyao Wu, Zhou Zhao, and Wei Zhou. 2019. Hierarchical human-like strategy for aspect-level sentiment classification with sentiment linguistic knowledge and reinforcement learning. *Neural Networks*.