Using Customer Service Dialogues for Satisfaction Analysis with Context-Assisted Multiple Instance Learning

Kaisong Song[1], Lidong Bing[1], Wei Gao[2], Jun Lin[1], Lujun Zhao[1]
Jiancheng Wang[3], Changlong Sun[1], Xiaozhong Liu[4], Qiong Zhang[1]
[1]Alibaba Group, Hangzhou, China
[2]Victoria University of Wellington, Wellington, New Zealand
[3]Soochow University, Suzhou, China
[4]Indiana University Bloomington, USA
{kaisong.sks,l.bing,linjun.lj,lujun.zlj,qz.zhang}@alibaba-inc.com,wei.gao@vuw.ac.nz
jiancheng.wang@qq.com, changlong.scl@taobao.com, liu237@indiana.edu

Abstract

Customers ask questions and customer service staffs answer their questions, which is the basic service model via multi-turn customer service (CS) dialogues on E-commerce platforms. Existing studies fail to provide comprehensive service satisfaction analysis, namely satisfaction polarity classification (e.g., well satisfied, met and unsatisfied) and sentimental utterance identification (e.g., positive, neutral and negative). In this paper, we conduct a pilot study on the task of service satisfaction analysis (SSA) based on multi-turn CS dialogues. We propose an extensible Context-Assisted Multiple Instance Learning (CAMIL) model to predict the sentiments of all the customer utterances and then aggregate those sentiments into service satisfaction polarity. After that, we propose a novel Context Clue Matching Mechanism (CCMM) to enhance the representations of all customer utterances with their matched context clues, i.e., sentiment and reasoning clues. We construct two CS dialogue datasets from a top E-commerce platform. Extensive experimental results are presented and contrasted against a few previous models to demonstrate the efficacy of our model. [1]

1 Introduction

In the past decades, E-commerce platforms, such as Amazon.com and Taobao.com[2], have evolved into the most comprehensive and prosperous business ecosystems. They not only deeply involve other traditional businesses such as payment and logistics, but also largely transform every aspect of retailing. Taking the customer service on Taobao as an example, third-party retailers are always online to answer any question at any stage of pre-sale, sale and after-sale, through an instant messenger within the platform. The topics of relevant customer service dialogues involve various aspects of online shopping, such as product information, return or exchange, logistics, etc. Based on a previous survey, over 77% of buyers on Taobao communicated with sellers before placing an order (Gao and Zhang, 2011). Therefore, such service dialogue data contain very important clues for sellers to improve their service quality.

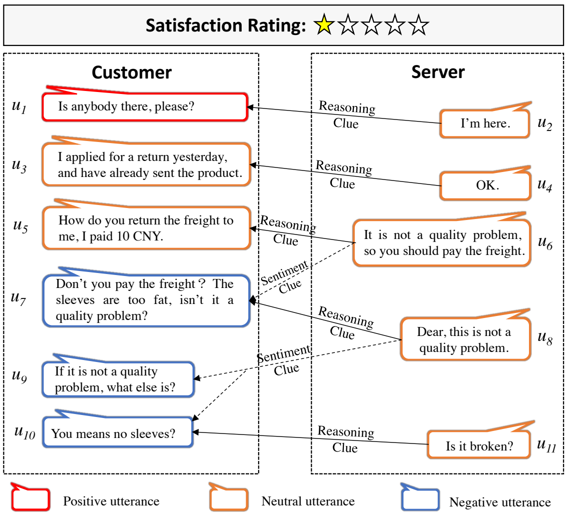
Satisfaction Rating: ★☆☆☆☆
Customer
Server
u_1 Is anybody there, please?
Reasoning Clue
I'm here. u_2
u_3 I applied for a return yesterday, and have already sent the product.
Reasoning Clue
OK. u_4
u_5 How do you return the freight to me, I paid 10 CNY.
Reasoning Clue
It is not a quality problem, so you should pay the freight. u_6
Sentiment Clue
u_7 Don't you pay the freight ? The sleeves are too fat, isn't it a quality problem?
Reasoning Clue
Dear, this is not a quality problem. u_8
Sentiment Clue
u_9 If it is not a quality problem, what else is?
u_{10} You means no sleeves?
Reasoning Clue
Is it broken? u_{11}
Positive utterance
Neutral utterance
Negative utterance

Figure 1: Customer utterance and context clues (i.e., sentiment and reasoning clues) alignments in multi-turn dialogue utterances of an *unsatisfied* customer service. The utterances (u_i) with *positive/neutral/negative* sentiments are denoted by *red*/*orange*/*blue* boxes.

Figure 1 depicts an exemplar dialogue of online customer service, which has a form of multi-turn dialogue between the customer and the customer service staff (or "the server" for short). In this dialogue, the customer is asking for refunding the freight he/she paid for sending back the product. At the end of service dialogue, the E-commerce platform invites the customer to score the service quality (e.g., using 1-5 stars denoting the extent of

[1] We have released the dataset at https://github.com/songkaisong/ssa.

[2] Taobao is a top E-commerce platform in China.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 198–207,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

satisfaction from "*very unsatisfied*" to "*very satisfied*") via instant messages or a grading interface. Evidently, the customer feels unsatisfied with the response. Automatically detecting such unsatisfactory service is important. For the retail shopkeepers, they can quickly locate such service dialogue and find out the reason to take remedial actions. For the platform, by detecting and analyzing such cases, the platform can define clear-cut rules, say "*not fitting well is not a quality issue, and the buyers should pay the freight for freight.*"

In this paper, we define a new task named **Service Satisfaction Analysis (SSA)**: *Given a service dialogue between the customer and the service staff, the task aims at predicting the customer's satisfaction, i.e., if the customer is satisfied by the responses from the service staff, meanwhile locating possible sentiment reasons, i.e., sentiment identification of the customer utterances*. For example, Figure 1 gives the satisfaction prediction of the service as "*unsatisfied*" and identifies the detailed sentiments of all customer utterances. Obviously, SSA focuses on two special cases of text classification over predefined satisfaction labels ("*well satisfied/met/unsatisfied*") and predefined sentiment labels ("*positive/neutral/negative*"). Text classification has been widely studied for decades, such as sentiment classification on product reviews (Song et al., 2017; Chen et al., 2017; Li et al., 2018, 2019), stance classification on tweets or blogs (Du et al., 2017; Liu, 2010), emotion classification for chit-chat (Majumder et al., 2018), etc. However, all these methods cannot deal with these two classification tasks simultaneously in a unified framework. Although recent studies on multi-task learning framework suggest that closely related tasks can improve each other mutually from separated supervision information (Ma et al., 2018; Cerisara et al., 2018; Wang et al., 2018b), the acquisition of sentence (or utterance)-level sentiment labels, which is required by multi-task learning, remains a laborious and expensive endeavor. In contrast, coarse-grained document (or dialogue)-level annotations are relatively easy to obtain due to the widespread use of opinion grading interfaces (e.g., ratings).

Recently, *Multiple Instance Learning* (MIL) framework is adopted for performing document-level and sentence-level sentiment classification simultaneously while only using document-level sentiment annotations (Zhou et al., 2009; Wang and Wan, 2018). However, these models are trained based on plain textual data which are in a much simpler form than our multi-turn dialogue structure. Specifically, customer service dialogue has unique characteristics. Customer utterances tend to have more sentiment changes during the customer service dialogue which affect customer's final satisfaction. Figure 1 illustrates that satisfaction polarity ("*unsatisfied*") is mostly embedded in the last few customer utterances (i.e., u_7, u_9 and u_{10})[3]. On the other hand, a well-trained server varies less by always expressing positive/neutral utterances which contain helpful sentiment clues and reasoning clues. In this work, both sentiment clue and reasoning clue are called context clues which can directly or indirectly influence satisfaction polarity and need to be given special treatments in the model.

To deal with the issues, we propose a novel and extensible Context-Assisted Multiple Instance Learning (CAMIL) model for the new SSA task, and utterance-level sentiment classification and dialogue-level satisfaction classification will be done simultaneously only under the supervision of satisfaction labels. We motivate the idea of our context-assisted modeling solution based on the hypothesis that if a customer utterance does not have enough information to create a sound vector representation for sentiment prediction, we try to enhance it with a complementary representation derived from context clues via our position-guided Context Clue Matching Mechanism (CCMM). Overall, our contributions are three-fold:

- We introduce a new SSA task based on customer service dialogues. We thus propose a novel CAMIL model to predict the sentiment distributions of all customer utterances, and then aggregate those distributions to determine the final satisfaction polarity.
- We further propose an automatic CCMM to associate each customer utterance with its most relevant context clues, and then generate a complementary vector which enhances the customer utterance representation for better sentiment classification to boost the final satisfaction classification.

[3] Given a specific customer utterance (u_7) in Figure 1, the server utterance (u_6) triggers the change of customer sentiment from "*neutral*" to "*negative*", and the server utterance (u_8) answers the question "*this is not a quality problem*".

- Two real-world CS dialogue datasets are collected from a top E-commerce platform. The experimental results demonstrate that our model is effective for the SSA task.

2 Related Work

Service satisfaction analysis (SSA) is closely related to sentiment analysis (SA), because the sentiment of the customer utterances is a basic clue signaling the customer's satisfaction. Existing SA works aim to predict sentiment polarities (*positive*, *neutral* and *negative*) for subjective texts in different granularities, such as word (Feng et al., 2015; Song et al., 2016), sentence (Ma et al., 2017), short text (Song et al., 2015) and document (Yang et al., 2018a). In these studies, subjective texts are always considered as a sequence of words[4]. More recently, some researchers started to explore the utterance-level structure for sentiment classification, such as modeling dialogues via a hierarchical RNN in both word level and utterance level (Cerisara et al., 2018) or keeping track of sentiment states of dialogue participants (Majumder et al., 2018). However, none of these works can do dialogue-level satisfaction classification and utterance-level sentiment classification simultaneously. Recent studies (Cerisara et al., 2018; Ma et al., 2018; Wang et al., 2018b) employing multi-task learning open a possibility to address this issue. However, these models must be trained under the supervision of both document-level and sentence-level sentiment labels in which the later are generally not easy to obtain.

Sentiment classification based on Multiple Instance Learning (MIL) frameworks (Wang and Wan, 2018; Angelidis and Lapata, 2018) aims to perform document-level and sentence-level sentiment classification tasks simultaneously with the supervision of document labels only. Angelidis and Lapata (2018) proposed an MIL model for fine-grained sentiment analysis. Wang and Wan (2018) further applied the model to peer-reviewed research papers by integrating a memory built from abstracts. However, their models are not suitable for our SSA task because they ignore the dialogue structure of arbitrary interactions between customers and servers. In contrast, we consider complex multi-turn interactions within dialogues and explore context clue matching between customer utterances and server utterances for multi-tasking in the SSA task. Specifically, we improve the basic MIL models by proposing a position-guided automatic context clue matching mechanism (CCMM) to conduct customer utterance and context clues alignments for better sentiment classification to boost satisfaction classification. Other related work related to sentiment analysis for subjective texts in different granularities include (Yang et al., 2016; Wu and Huang, 2016; Yang et al., 2018b; Du et al., 2017; Wang et al., 2018a; Song et al., 2019).

3 Context-Assisted MIL Network

In order to predict service satisfaction and identify sentiments of all customer utterances with available satisfaction labels, we propose a CAMIL model based on multiple instance learning approach. Figure 2 shows the architecture of our model which consists of three layers: *Input Representation Layer*, *Sentiment Classification Layer* and *Satisfaction Classification Layer*. In this section, we will describe the model in detail.

3.1 Input Representation Layer

Let each utterance $u_i = \left[w_1, ..., w_{|u_i|}\right]$ be a sequence of words. By adopting word embeddings and semantic composition models such as Recurrent Neural Network (RNN), we can learn the utterance representation. In this work, we adopt a standard LSTM model (Hochreiter and Schmidhuber, 1997) to learn a fixed-size utterance representation $\boldsymbol{v}_{u_i} \in \mathbb{R}^k$, where k is the size of LSTM hidden state. Specifically, we first convert the words in each utterance u_i to the corresponding word embeddings $\boldsymbol{E}_{u_i} \in \mathbb{R}^{d\times|u_i|}$ which are then fed into a LSTM for obtaining the last hidden state as the $\boldsymbol{v}_{u_i}$, where d is the dimensionality of word embeddings. Formally, we have $\boldsymbol{v}_{u_i} = \mathrm{LSTM}(\boldsymbol{E}_{u_i})$.

We conjecture that the participants (i.e., customer and server) play different roles in CS dialogue. Our hypothesis is that satisfaction polarity can be more or less conveyed by the sentiments of key customer utterances, and meanwhile the sentiments of server utterances are generally polite or non-negative and contain text with context clues which complement the target customer's ut-

[4] From plain sentiment analysis viewpoint, sentence and utterance are treated the same since a dialogue is considered as a chunk of plain texts and the matching between utterances is ignored. So, an utterance is seen as a sentence and a dialogue as a document. Here, we use sentence and utterance interchangeably when it comes to plain sentiment analysis models.

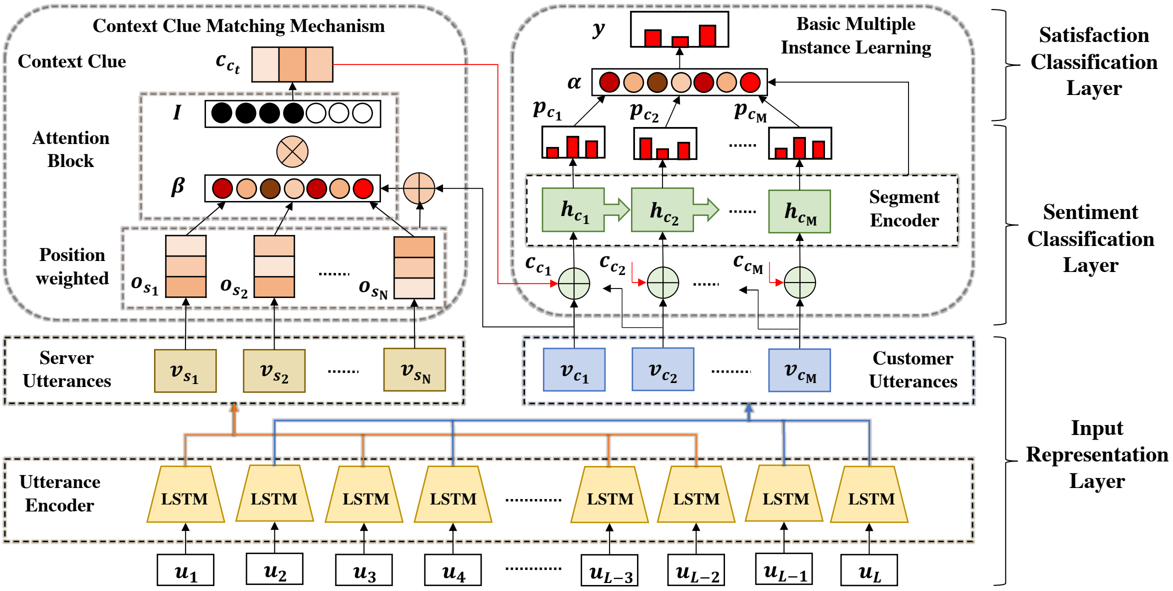

Figure 2: The architecture of our Context-Assisted Multiple Instance Learning (CAMIL) network. The model consists of four modules: Input Representation Layer, Sentiment Classification Layer, Satisfaction Classification Layer and Context Clue Matching Mechanism.

terances and indirectly affect satisfaction polarity. Thus, we separately denote the customer utterance representations as $\{\boldsymbol{v}_{c_1}, \boldsymbol{v}_{c_2}, ..., \boldsymbol{v}_{c_M}\}$ and server utterance representations as $\{\boldsymbol{v}_{s_1}, \boldsymbol{v}_{s_2}, ..., \boldsymbol{v}_{s_N}\}$, where $M + N = L$ and L is the total number of utterances in the dialogue.

3.2 Sentiment Classification Layer

Customer utterances tend to have more direct impact on the dominating satisfaction polarity. However, short utterance texts may not contain enough information for semantic representation. Thus, considering context to enhance utterance representation is a natural and reasonable choice. Given a specific customer utterance vector $\boldsymbol{v}_{c_t}$, we use a context clue matching mechanism, namely CCMM (see Section 4), to produce matched context representation $\boldsymbol{c}_{c_t} \in \mathbb{R}^k$ as below:

$$\boldsymbol{c}_{c_t} = \text{CCMM}\left(\boldsymbol{v}_{c_t}, \{\boldsymbol{v}_{s_{t'}} | 1 \leq t' \leq N\}\right) \quad (1)$$

where $\boldsymbol{v}_{s_{t'}}$ is any server utterance representation.

After that, $\boldsymbol{v}_{c_t}$ can be enhanced by $\boldsymbol{c}_{c_t}$ via concatenation for a combined representation $\bar{\boldsymbol{v}}_{c_t} = \boldsymbol{v}_{c_t} \oplus \boldsymbol{c}_{c_t}$. Compared to $\boldsymbol{v}_{c_t}$, $\bar{\boldsymbol{v}}_{c_t} \in \mathbb{R}^{2k}$ contains more evidence for sentiment prediction. Then, we feed the representation sequence $\{\bar{\boldsymbol{v}}_{c_1}, \bar{\boldsymbol{v}}_{c_2}, ..., \bar{\boldsymbol{v}}_{c_M}\}$ into a standard LSTM for obtaining a segment representation $\boldsymbol{h}_{c_t} \in \mathbb{R}^k$ at each time step t, i.e., $\boldsymbol{h}_{c_t} = \text{LSTM}\left(\bar{\boldsymbol{v}}_{c_t}\right)$.

Finally, each segment representation $\boldsymbol{h}_{c_t}$ is fed into a linear layer and then a softmax function for predicting its sentiment distribution over sentiment labels $\mathcal{G} = \{$*positive, neutral, negative*$\}$:

$$\boldsymbol{p}_{c_t} = \text{softmax}(\mathbf{W}^s \boldsymbol{h}_{c_t} + \boldsymbol{b}^s) \quad (2)$$

where $\mathbf{W}^s \in \mathbb{R}^{|\mathcal{G}| \times k}$ and $\boldsymbol{b}^s \in \mathbb{R}^{|\mathcal{G}|}$ are trainable parameters shared across all segments, and $\boldsymbol{p}_{c_t} \in \mathbb{R}^{|\mathcal{G}|}$ is the sentiment distribution for utterance u_{c_t}.

3.3 Satisfaction Classification Layer

In the simplest case, satisfaction polarities $\mathcal{C} = \{$*well satisfied, met, unsatisfied*$\}$ can be computed by averaging all predicted sentiment distributions of customer utterances as $\boldsymbol{y} = \frac{1}{M} \sum_{t \in [1,M]} \boldsymbol{p}_{c_t}$. However, it is a crude way of combining sentiment distributions uniformly, as not all distributions convey equally important sentiment clues. In Figure 1, for example, the satisfaction polarity ("unsatisfied") is mostly determined by customer utterances u_7, u_9 and u_{10} which are relatively more crucial than other ones. We opt for an attention mechanism to reward segments that are more likely to be good sentiment predictors. Therefore, we measure the importance of each segment representation $\boldsymbol{h}_{c_t}$ through a scoring function using feed forward neural network as below:

$$\alpha_{c_t} = \text{softmax}\left(\boldsymbol{v}^{\text{T}} \tanh(\boldsymbol{W}^u \boldsymbol{h}_{c_t} + \boldsymbol{b}^u)\right) \quad (3)$$

where $\mathbf{W}^u \in \mathbb{R}^{k\times k}$, $\boldsymbol{b}^u \in \mathbb{R}^k$ and $\boldsymbol{v} \in \mathbb{R}^k$ are trainable parameters, $\boldsymbol{v}$ can be seen as a high-level representation of a fixed query "*what is the informative segment*" like that used in (Yang et al., 2016). Finally, we obtain the satisfaction distribution $\boldsymbol{y} \in \mathbb{R}^{|\mathcal{C}|}$ as the weighted sum of sentiment distributions of all the customer utterances by:

$$\boldsymbol{y} = \sum_{t\in[1,\mathrm{M}]} \alpha_{c_t}\boldsymbol{p}_{c_t} \tag{4}$$

3.4 Training and Parameter Learning

Note that in the training dataset, our approach only needs the dialogue's satisfaction labels while the utterance sentiment labels are unobserved. Therefore, we use the categorical cross-entropy loss to minimize the error between the distribution of the output satisfaction polarity and that of the gold-standard satisfaction label of the dialogue by:

$$\mathcal{L}(\Theta) = -\sum_{j\in[1,T]}\sum_{i\in\mathcal{C}} g_i^j \log(y_i^j) \tag{5}$$

where g_i^j is 1 or 0 indicating whether the i^{th} class is a correct answer for the j^{th} training instance, y_i^j is the predicted satisfaction probability distribution, and Θ denotes the trainable parameter set.

After learning Θ, we feed each test instance into the final model, and the label with the highest probability stands for the predicted satisfaction polarity. We use back propagation to calculate the gradients of all the model parameters, and update them with Momentum optimizer (Qian, 1999).

4 Context Clue Matching Mechanism

Server utterances provide helpful context clues which can be defined as sentiment and reasoning clues by the positions of server utterances. Thus, we introduce the position-guided automatic context clue matching mechanism (CCMM) used to match each customer utterance with its most related server utterances, which contain two layers: the *position attention layer* and the *utterance attention layer*.

Sentiment and Reasoning Clues: Server utterances provide helpful context clues for each targeted customer utterance. Here, we aim to locate helpful context clues in server utterances which are categorized as sentiment clues and reasoning clues. Sentiment clues refer to the server utterances that appear preceding the targeted customer utterance and trigger its sentiment expression, such as server utterance u_6 leading to customer displeasure of the utterance u_7 in the Figure 1. Reasoning clues are the server utterances that appear following the targeted customer utterance and respond to its concerns, such as server utterance u_6 responding to the customer utterance u_5 in the Figure 1. Both types of clues are identified by the proposed attention layers along with position information.

Position Attention Layer: Typically, customer utterances are more likely to be triggered or answered by the server utterances near them. Let $p(\cdot)$ denote the position function of any utterance in the original dialogue, such as $p(u_{c_2}) = 3$ in Figure 1. For any customer utterance u_{c_t}, the preceding server utterances $\{u_{s_{t'}}|p(u_{s_{t'}}) < p(u_{c_t})\}$ may provide sentiment clues, and the following server utterances $\{u_{s_{t'}}|p(u_{s_{t'}}) > p(u_{c_t})\}$ may contain reasoning clues. By considering both directions, we compute the position attention weight $g(\cdot)$ by:

$$g(u_{c_t}, u_{s_{t'}}) = 1 - \frac{|p(u_{c_t}) - p(u_{s_{t'}})|}{L} \tag{6}$$

Then, the weighted output after this layer is formulated as below:

$$\boldsymbol{o}_{s_{t'}} = g(u_{c_t}, u_{s_{t'}}) * \boldsymbol{h}_{s_{t'}} * I(u_{c_t}, u_{s_{t'}}) \tag{7}$$

where $\boldsymbol{o}_{s_{t'}} \in \mathbb{R}^k$ is the weighted $\boldsymbol{h}_{s_{t'}}$ for $t' \in [1, N]$, the notation $I(\cdot)$ denotes a masking function which can be used to reserve either only sentiment clues (i.e., if $p(u_{c_t}) > p(u_{s_{t'}})$, $I(u_{c_t}, u_{s_{t'}})$ equals to 1, or 0 otherwise) or only reasoning clues (i.e., if $p(u_{c_t}) < p(u_{s_{t'}})$, $I(u_{c_t}, u_{s_{t'}})$ equals to 1, or 0 otherwise). Here, we suggest to consider both sentiment and reasoning clues, so $I(u_{c_t}, u_{s_{t'}})$ is a constant 1. Finally, we construct memory $\boldsymbol{O} \in \mathbb{R}^{k\times N}$ as below:

$$\boldsymbol{O} = [\boldsymbol{o}_{s_1}; \boldsymbol{o}_{s_2}; \ldots; \boldsymbol{o}_{s_\mathrm{N}}] \tag{8}$$

Utterance Attention Layer: Only a fraction of server utterances can match every customer utterance in sentiment or content, such as the exemplar dialogue in Figure 1. So, we introduce an attention strategy which enables our model to attend on server utterances of different importance when constructing a complementary context representation for any customer utterance. Considering customer utterance representation $\boldsymbol{h}_{c_t}$ as an index, we can produce a context vector $\boldsymbol{c}_{c_t} \in \mathbb{R}^k$ using a weighted sum of each piece $\boldsymbol{o}_{s_{t'}}$ of memory $\mathbf{O}$:

$$\boldsymbol{c}_{c_t} = \sum_{t'\in[1,N]} \beta_{s_{t'}}\boldsymbol{o}_{s_{t'}} \tag{9}$$

Statistics items	Clothes	Makeup
# Dialogues	10,000	3,540
# US (unsatisfied)	2,302	1,180
# MT (met)	6,399	1,180
# WS (well satisfied)	1,299	1,180
Avg# Utterances	25.99	26.67
Avg# Words	7.11	7.32
# Segments	123,242	46,255
# NG (negative)	12,619	6,130
# NE (neutral)	97,380	33,158
# PO (positive)	13,243	6,976

Table 1: Statistics of the datasets we collected.

where $\beta_{s_{t'}} \in [0,1]$ is the attention weight calculated based on a scoring function using a feed forward neural network as follow:

$$\beta_{s_{t'}} = \mathrm{softmax}\left(\bar{\mathbf{v}}^{\mathrm{T}} \tanh(\mathbf{W}^c \begin{bmatrix} \boldsymbol{o}_{s_{t'}} \\ \boldsymbol{h}_{ct} \end{bmatrix} + \boldsymbol{b}^c)\right) \tag{10}$$

where $\mathbf{W}^c \in \mathbb{R}^{2k\times 2k}$, $\bar{\mathbf{v}} \in \mathbb{R}^{2k}$ and $\boldsymbol{b}^c \in \mathbb{R}^{2k}$ are trainable parameters.

5 Experiments and Results

5.1 Dataset and Experimental Settings

Our experiments are conducted based on two Chinese CS dialogue datasets, namely **Clothes** and **Makeup**, collected from a top E-commerce platform. Note that our proposed method is language independent and can be applied to other languages directly. **Clothes** is a corpus with 10K dialogues in the *Clothes* domain and **Makeup** is a balanced corpus with 3,540 dialogues in the *Makeup* domain. Both datasets have service satisfaction ratings in 1-5 stars from customer feedbacks. Meanwhile, we also annotate all the utterances in both datasets with sentiment labels for testing. In this study, we conduct two classification tasks: one is to predict in three satisfaction classes, i.e., "*unsatisfied*" (1-2 stars), "*met*" (3 stars) and "*satisfied*" (4-5 stars), and the other is to predict in three sentiment classes, i.e., "*negative/neutral/positive*". All texts are tokenized by a popular Chinese word segmentation utility called jieba[5]. After preprocessing, the datasets are partitioned for training, validation and test with a 80/10/10 split. A summary of statistics for both datasets are given in Table 1.

For all the methods, we apply fine-tuning for the word vectors, which can improve the performance. The word vectors are initialized by word embeddings that are trained on both datasets with CBOW (Mikolov et al., 2013), where the dimension is 300 and the vocabulary size is 23.3K. Other trainable model parameters are initialized by sampling values from a uniform distribution $\mathcal{U}(-0.01, 0.01)$. The size of LSTM hidden states k is set as 128. The hyper-parameters are tuned on the validation set. Specifically, the initial learning rate is fixed as 0.1, the dropout rate is 0.2, the batch size is 32 and the number of epochs is 20. The performances of both satisfaction and sentiment classifications are evaluated using standard *Macro F1* and *Accuracy*.

5.2 Comparative Study

We compare our proposed approach with the following state-of-the-art Sentiment Analysis (SA) methods which can be grouped into two types: plain SA models and dialogue SA models.

Plain SA models consider dialogue as plain text and ignore utterance matching, say, utterance is seen as sentence and dialogue as document.

1) **LSTM**: We use word vectors as the input of a standard LSTM (Hochreiter and Schmidhuber, 1997) and feed the last hidden state into a softmax layer for satisfaction prediction.

2) **HAN**: A hierarchical attention network for document classification (Yang et al., 2016), which has two levels of attention mechanisms applied at word- and utterance-level, enabling it to attend differentially to more and less important content when constructing the dialogue representation and feeding it into a softmax layer for classification.

3) **HRN**: A hierarchical recurrent network for joint sentiment and act sequence recognition (Cerisara et al., 2018). It uses a bi-directional LSTM to represent utterances which are then fed into a standard LSTM for dialogue representation as the input of a softmax layer for classification.

4) **MILNET**: A multiple instance learning network for document-level and sentence-level sentiment analysis (Angelidis and Lapata, 2018). The original method is designed for plain textual data, which does not consider CS dialogue structure. In addition, their method ignores long-range dependencies among customer sentiments (i.e., without segment encoder in Figure 2).

Dialogue SA models consider utterance matching between customer and server utterances.

5) **HMN**: A hierarchical matching network for sentiment analysis, which uses a question-answer bidirectional matching layer to learn the matching

[5] https://pypi.org/project/jieba/

vector of each QA pair (i.e., customer utterance, server utterance) and then characterizes the importance of the generated matching vectors via a self-matching attention layer (Shen et al., 2018). However, the amount of pairs within a dialogue is huge, which leads to expensive calculations. Meanwhile, it considers the sentiments of server utterances, which will mislead final prediction.

6) **CAMIL$_s$**, **CAMIL$_r$** and **CAMIL$_{full}$**: Our CAMIL models with only sentiment clues, only reasoning clues, and both of them, respectively, by setting masking function (see Equation 7).

All the methods are implemented by ourselves with TensorFlow[6] and run on a server configured with a Tesla V100 GPU, 2 CPU and 32G memory.

Results and Analysis: The results of comparisons are reported in Table 2. It indicates that LSTM cannot compete with other methods because it simply considers dialogues as word sequences but ignores the utterance matching. HAN and HRN perform much better by using a two-layer architecture (i.e., utterance and dialogue), but they ignore the utterance interactions. Besides, HRN treats the sentiment analysis task and the service satisfaction analysis task separately, and ignores their sentiment dependence. HMN uses a heuristic question-answering matching strategy, which is not enough flexible and easily causes mismatching issues. MILNET is the most related work, but its simplistic alignment model weakens prediction performance when facing on our complex customer service dialogue structure. MILNET however does not consider the dialogue structure and introduces unrelated sentiments from server utterances. CAMIL$_r$ and CAMIL$_s$ only consider either sentiment or reasoning clues, so they cannot compete with CAMIL$_{full}$ which considers both in dialogues. Partially configured model CAMIL_r (or CAMIL_s) only considers sentiment (or reasoning) clues and performs worse than our full model CAMIL. This verifies that both types of clues are helpful and complementary, and they should be employed simultaneously.

On Clothes corpus, compared to the *met* class, the performances of all models on the *satisfied* class are much worse, because when the two classes cannot be well distinguished the models tend to predict the majority class (i.e., *met*) to minimize the loss. On Makeup corpus which is a balanced dataset, the performances on the *met* and *satisfied* classes are less distinctive, but both are consistently worse than the *unsatisfied* class.

Methods	**Clothes**				
	WS F1	MT F1	US F1	MacroF1	Acc.
LSTM	0.264	0.772	0.634	0.557	0.684
HAN	0.515	0.817	0.704	0.679	0.755
HRN	0.508	0.835	0.676	0.673	0.766
MILNET	0.382	0.823	0.708	0.638	0.753
HMN	0.441	0.833	0.696	0.657	0.763
CAMIL$_s$	0.450	0.822	0.713	0.665	0.767
CAMIL$_r$	0.448	0.827	0.704	0.659	0.764
CAMIL$_{full}$	**0.554**	**0.844**	**0.715**	**0.704**	**0.783**
Methods	**Makeup**				
	WS F1	MT F1	US F1	MacroF1	Acc.
LSTM	0.585	0.597	0.724	0.635	0.632
HAN	0.684	0.713	0.848	0.748	0.748
HRN	0.720	0.686	0.858	0.755	0.754
MILNET	0.720	0.689	0.849	0.753	0.751
HMN	0.735	0.731	0.834	0.766	0.768
CAMIL$_s$	0.693	0.710	0.840	0.747	0.748
CAMIL$_r$	0.687	0.705	0.861	0.751	0.754
CAMIL$_{full}$	**0.738**	**0.745**	**0.874**	**0.786**	**0.785**

Table 2: Results of different satisfaction classification methods. The best results are highlighted.

Methods	**Clothes**				
	WS F1	MT F1	US F1	MacroF1	Acc.
Server	0.346	0.785	0.589	0.573	0.689
Customer	0.553	0.824	0.663	0.681	0.759
NoPos	0.554	0.838	0.673	0.688	0.771
Average	0.070	0.789	0.347	0.402	0.671
Voting	0.059	0.776	0.046	0.293	0.633
CAMIL$_{full}$	**0.554**	**0.844**	**0.715**	**0.704**	**0.783**
Methods	**Makeup**				
	WS F1	MT F1	US F1	MacroF1	Acc.
Server	0.597	0.630	0.745	0.657	0.655
Customer	0.735	0.687	0.790	0.737	0.734
NoPos	0.731	0.742	0.864	0.779	0.779
Average	0.578	0.708	0.834	0.706	0.714
Voting	0.231	0.016	0.553	0.267	0.387
CAMIL$_{full}$	**0.738**	**0.745**	**0.874**	**0.786**	**0.785**

Table 3: Results of different model configurations.

5.3 Ablation Study

Different model configurations can largely affect the performance. We implement several model variants for ablation tests: **Server** and **Customer** consider only server and customer utterances in a dialogue, respectively. **NoPos** ignores the prior position information. **Average** takes the average of all the sentiment distributions for classification. **Voting** directly maps the majority sentiment into satisfaction prediction, i.e., negative → unsatisfied, neutral → met, positive → well-satisfied. The results of comparisons are reported in Table 3.

In Table 3, we can observe that **Customer** outperforms **Server** by a large margin, which indicates that service satisfaction is mostly related to

[6] https://www.tensorflow.org/

Dataset	Class	positive	neutral	negative
Clothes	WS	0.278	0.712	0.009
	MT	0.020	0.893	0.086
	US	0.042	0.513	0.443
Makeup	WS	0.273	0.718	0.009
	MT	0.012	0.893	0.094
	US	0.046	0.598	0.355

Table 4: Sentiment distribution in satisfaction classes.

Methods	Clothes				
	PO F1	NE F1	NG F1	MacroF1	Acc.
MILNET	0.441	0.814	0.404	0.553	0.713
CAMIL$_s$	**0.545**	0.867	0.506	0.639	0.787
CAMIL$_r$	0.470	0.870	0.529	0.623	0.792
CAMIL$_{full}$	0.484	**0.893**	**0.555**	**0.644**	**0.824**
Methods	**Makeup**				
	PO F1	NE F1	NG F1	MacroF1	Acc.
MILNET	0.447	0.387	0.416	0.417	0.410
CAMIL$_s$	**0.566**	0.672	0.501	0.580	0.609
CAMIL$_r$	0.556	0.600	0.488	0.548	0.561
CAMIL$_{full}$	0.544	**0.725**	**0.516**	**0.595**	**0.647**

Table 5: Results of sentiment classification by different models. The best results are highlighted.

$\mathbf{C}_1$: I have been waiting too long time! Negative $[NG;NE]$
$\mathbf{S}_1$: Dear, I am the customer service agent. May I help you?
$\mathbf{C}_2$: I haven't received the goods yet! Why? Negative $[NG;NG]$
$\mathbf{S}_2$: Dear, I'm reading your messages.
$\mathbf{C}_3$: As a rule, it should be received today. Neutral $[NE;NE]$
$\mathbf{S}_3$: Dear, I feel very sorry about it.
$\mathbf{S}_4$: I will help you push the workers.
$\mathbf{C}_4$: You deserve bad ratings! Negative $[NG;NG]$
$\mathbf{C}_5$: It is supposed to be shipped here in 3 days. Many days have passed! Negative $[NG;NE]$
$\mathbf{S}_5$: Dear, I have already pushed them.
$\mathbf{C}_6$: Tell me when the goods will arrive? Negative $[NG;NE]$
$\mathbf{C}_7$: I will apply for refund if I don't get it tomorrow. Neutral $[NE;NE]$
$\mathbf{S}_6$: I have urged them.
(*A few hours passed.*)
$\mathbf{C}_8$: I don't want it. I will apply for refund, immediately! Negative $[NG;NG]$
$\mathbf{S}_7$: I'm sorry for the inconvenience.
$\mathbf{S}_8$: Please wait for a few more days.
$\mathbf{S}_9$: I have pushed them twice!
$\mathbf{C}_9$: However, I know this is useless. Negative $[NG;NG]$

Satisfaction: Unsatisfied $[US;MT]$

Figure 3: An example dialogue with predictions. C$_i$ (S$_j$) are customer (server) utterances. True labels are underlined. The predictions by our model and MILNET are colored in *red* and *blue* in the brackets, respectively.

the sentiments embedded in the customer utterances. However, its performance is still lower than CAMIL$_{full}$, suggesting that server utterances can provide helpful context clues. NoPos performs well but worse than CAMIL$_{full}$ since the position information provides prior knowledge for guiding context clue matching. Average and Voting are sub-optimal choices because not all the sentiment distributions contribute equally to the satisfaction polarity and the majority sentiment polarity also does not correlate strongly with it.

5.4 Results on Sentiment Classification

Table 4 shows another statistics of our datasets, i.e., the distribution of sentiment labels over each service satisfaction polarity, which reflects the imbalanced situation of utterance-level sentiments in real customer service dialogues.

In Table 5, we compare the sentiment prediction results of MILNET, CAMIL$_r$, CAMIL$_s$ and CAMIL$_{full}$. CAMIL$_r$ and CAMIL$_s$ perform worse than CAMIL$_{full}$ because they only consider partial context information. CAMIL$_{full}$ is the best mainly due to its accurate context clue matching. Thus, our proposed approach is more adaptive to the service satisfaction analysis task based on the customer service dialogues.

5.5 Case Study

Figure 1 illustrates our prediction results with an example dialogue which is translated from Chinese text. For brevity, we use **C/S** to denote customer/server utterance. Our model predicts the label "*unsatisfied*" correctly and also predicts reasonable sentiment polarities for customer utterances. Considering the context, customer utterances $\mathbf{C}_{1,5,6}$ are "*negative*" but predicted as "*neutral*" by MILNET because MILNET predicts sentiments only from target utterance itself and ignores context information. In addition, the sentiments of the customer utterances $\mathbf{C}_{4,5}$ and $\mathbf{C}_9$ tend to have larger influences on deciding the satisfaction polarity because $\mathbf{C}_4$ clearly conveys "*unsatisfied*" attitude, $\mathbf{C}_5$ complains about delay and $\mathbf{C}_9$ criticizes the low service quality.

We also visualize the attention weights in Figure 4 to explain our prediction results. For each customer utterance $\mathbf{C}_i$, we give the attention weights $\beta_{S_{t'}}$ on all the server utterances (see Formula 10). Furthermore, we also visualize the attention rates α_{c_t} on the customer utterances (see Formula 3). *Lighter colors denote smaller values.* From Figure 4, we can see that the customer utterances $\mathbf{C}_{4,5,9}$ have higher attention weights because customer attitudes are intuitively formed at the end of the dialogues (i.e., $\mathbf{C}_9$) or determined by explicit sentiments (i.e., $\mathbf{C}_4$). In this example, the customer is finally unhappy with the provided solution, and the sentiments did not change through the whole dialogue. We can also see that customer

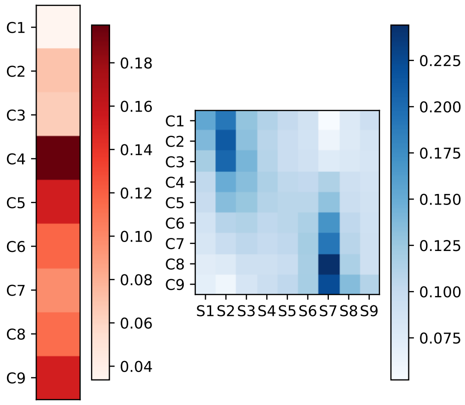

Figure 4: The visualization of attention rates α_{c_t} (Left in *red*) and $\beta_{s_{t'}}$ (Right in *blue*) for the given example.

utterances are influenced by server utterances. For example, $\mathbf{C}_{1-3}$ are related to $\mathbf{S}_2$, $\mathbf{C}_{4,5}$ are related to $\mathbf{S}_{2,3,7}$, and $\mathbf{C}_{6-9}$ are related to $\mathbf{S}_7$. This again validates the fact that customer utterances are related to the server utterances near them. Meanwhile, customer utterances may provide different types of context clues (i.e., sentiment and reasoning). For a specific server utterance $\mathbf{S}_7$, it provides explicit sentiment clue for $\mathbf{C}_9$ and also gives reasoning clue for $\mathbf{C}_8$.

In-depth Analysis: CAMIL$_{full}$ is only trained based on satisfaction labels, thus the laborious acquisition of sentiment labels is unnecessary. However, we would point out that lack of sentiment labels will inevitably lead to difficulties on identifying positive/negative utterances from those neutral ones. We will study to alleviate it in the future.

Our general observation is that the sentiment of customers at the beginning cannot largely determine the service satisfaction at the end. This is because the sentiment of the customers can vary with different quality of service during the dialogue, and the final service satisfaction results from the overall sentiments of important customer utterances in the dialogue (see the attention weights in Figure 4). To verify this, we design a heuristic baseline called Mapping which directly maps the initial *negative*, *neutral* and *positive* sentiment of customer to the corresponding service satisfaction, i.e., *unsatisfied*, *met* and *satisfied*. The satisfaction classification results are displayed in the Table 6.

In Table 6, we can observe that the Mapping method is far worse than our model. One reason is that the service dialogues in our datasets have more than 25 utterances in average (See the statistics in Table 1) and contain a large proportion of complex interactions. Besides, the sentiment change is closely related to the quality of service and it is very common in our datasets. Thus, using such simple correlation does not work well in our complex dialogue scenarios.

Methods	Clothes		Makeup	
	MacroF1	Acc.	MacroF1	Acc.
Mapping	0.625	0.487	0.471	0.430
CAMIL$_{full}$	**0.704**	**0.783**	**0.786**	**0.785**

Table 6: Satisfaction classification comparison between our method and a heuristic mapping method.

6 Conclusions and Future Work

In this paper, we propose a novel CAMIL model for the SSA task. We first propose a basic MIL approach with the inputs of context-matched customer utterances, then predict the utterance-level sentiment polarities and dialogue-level satisfaction polarities simultaneously. In addition, we propose a context clue matching mechanism (CCMM) to match any customer utterance with the most related server utterances. Experimental results on two real-world datasets indicate our method clearly outperforms some state-of-the-art baseline models on the two SSA subtasks, i.e., service satisfaction polarity classification and utterance sentiment classification, which are performed simultaneously. We have made our datasets publicly available.

In the future, we will further improve our method by learning the correlation between the customer utterances and the server utterances. In addition, we will study other interesting tasks in customer service dialogues, such as outcome prediction or opinion change.

7 Acknowledgements

We thank all the reviewers for their insightful comments and helpful suggestions. This work is supported by National Key R&D Program of China (2018YFC0830200; 2018YFC0830206).

References

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *TACL*, 6:17–31.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of COLING*, pages 745–754.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of EMNLP*, pages 452–461.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of IJCAI*, pages 3988–3994.

Shi Feng, Kaisong Song, Daling Wang, and Ge Yu. 2015. A word-emoticon mutual reinforcement ranking model for building sentiment lexicon from massive collection of microblogs. *World Wide Web*, 18(4):949–967.

Jie Gao and Zhenghua Zhang. 2011. User satisfaction of ali wangwang, an instant messenger tool. In *Theory, Methods, Tools and Practice - DUXU 2011, Part of HCI*, pages 414–420.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of ACL*, pages 946–956.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of AAAI*, pages 6714–6721.

Bing Liu. 2010. Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition.*, pages 627–666.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of IJCAI*, pages 4068–4074.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Proceedings of WWW*, pages 585–593.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2018. Dialoguernn: An attentive RNN for emotion detection in conversations. *CoRR*, abs/1811.00405.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR Workshop*.

Ning Qian. 1999. On the momentum term in gradient descent learning algorithms. *Neural Networks*, 12(1):145–151.

Chenlin Shen, Changlong Sun, Jingjing Wang, Yangyang Kang, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2018. Sentiment classification towards question-answering with hierarchical matching network. In *Proceedings of EMNLP*, pages 3654–3663.

Kaisong Song, Shi Feng, Wei Gao, Daling Wang, Ge Yu, and Kam-Fai Wong. 2015. Personalized sentiment classification based on latent individuality of microblog users. In *Proceedings of IJCAI*, pages 2277–2283.

Kaisong Song, Wei Gao, Ling Chen, Shi Feng, Daling Wang, and Chengqi Zhang. 2016. Build emotion lexicon from the mood of crowd via topic-assisted joint non-negative matrix factorization. In *Proceedings of SIGIR*, pages 773–776.

Kaisong Song, Wei Gao, Shi Feng, Daling Wang, Kam-Fai Wong, and Chengqi Zhang. 2017. Recommendation vs sentiment analysis: A text-driven latent factor model for rating prediction with cold-start awareness. In *Proceedings of IJCAI*, pages 2744–2750.

Kaisong Song, Wei Gao, Lujun Zhao, Jun Lin, Changlong Sun, and Xiaozhong Liu. 2019. Cold-start aware deep memory network for multi-entity aspect-based sentiment analysis. In *Proceedings of IJCAI*, pages 5197–5203.

Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. 2018a. Aspect sentiment classification with both word-level and clause-level attention networks. In *Proceedings of IJCAI*, pages 4439–4445.

Ke Wang and Xiaojun Wan. 2018. Sentiment analysis of peer review texts for scholarly papers. In *Proceedings of SIGIR*, pages 175–184.

Weichao Wang, Shi Feng, Wei Gao, Daling Wang, and Yifei Zhang. 2018b. Personalized microblog sentiment classification via adversarial cross-lingual multi-task learning. In *Proceedings of EMNLP*, pages 338–348.

Fangzhao Wu and Yongfeng Huang. 2016. Personalized microblog sentiment classification via multi-task learning. In *Proceedings of AAAI*, pages 3059–3065.

Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. 2018a. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *Proceedings of AAAI*, pages 6029–6036.

Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. 2018b. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *Proceedings of AAAI*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL*, pages 1480–1489.

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li. 2009. Multi-instance learning by treating instances as non-i.i.d. samples. In *Proceedings of ICML*, pages 1249–1256.