CUSTOMER SATISFACTION ESTIMATION USING UNSUPERVISED REPRESENTATION LEARNING WITH MULTI-FORMAT PREDICTION LOSS

Atsushi Ando, Yumiko Murata, Ryo Masumura, Satoshi Suzuki, Naoki Makishima,
Takafumi Moriya, Takanori Ashihara, Hiroshi Sato

NTT Corporation, Japan

ABSTRACT

We propose a new Customer Satisfaction Estimation (CSE) method that utilizes unsupervised representation learning. Though conventional methods have improved both the heuristic features and the estimation models, their performance is still insufficient as only small amounts of labeled training data can be expected. To mitigate this problem, the proposed method leverages a large amount of unlabeled data by unsupervised representation learning based on self-training. The key advance of the proposed method is to introduce a Multi-Format Prediction (MFP) loss to improve the performance of self-training for the inputs that contain both continuous and biased discrete features such as the number of occurrences of a particular word. MFP loss uses two loss functions based on regression and weighted binary classification to reconstruct both types of features with high accuracy. Experiments on real English contact center calls reveal the improved CSE performance attained by the proposed method.

Index Terms— Customer Satisfaction Estimation, Unsupervised Representation Learning, Multi-Format Prediction Loss

1. INTRODUCTION

Customer Satisfaction Estimation (CSE) in a contact center call is one of the most critical tasks in speech emotion recognition. It has many applications such as improving the efficiency of voice-of-customer analysis [1] and automatic evaluation of agents [2]. The tasks of CSE fall into two categories: call-level and turn-level CSE. Call-level CSE provides an estimate of the customer's satisfaction with the overall call [3]. Turn-level CSE, on the other hand, estimates the satisfaction value of each customer turn during a call [4]. A turn is defined as a series of utterance segments until the speaker changes. This paper aims to solve both categories of CSEs to support a greater variety of applications.

Various methods have been investigated for CSE. One of the major approaches is based on feature engineering and several types of heuristic features have been developed. Acoustic features such as statistics of fundamental frequency, power, and duration in customer utterances are widely used [5, 6]. Lexical features such as word N-gram or Bag-of-Words (BoW) have also been reported to be effective [7]. Some studies have proposed interactive features like turn overlap or call dominance [8, 9] and dialog event features such as answer repetition [10]. In addition, most of the recent work combines heuristic features with Deep Neural Networks (DNNs)-based recognition models. They can learn manifold cues and contextual information of customer satisfaction [11–14]. In our recent work [15, 16], we introduced a new DNN-based model, called the Hierarchical Multi-Task (HMT) model, to leverage the relationship between call-level and turn-level customer satisfaction degrees.

Despite these improvements in both feature engineering and model architecture, CSE is still challenging. One difficulty is the data limitations. It is expensive to collect a large amount of training data with customer satisfaction labels because it requires multiple professional annotators, such as supervisors, to get stable ground truths. Although several studies employ self-reported customer satisfaction labels to collect a large amount of labeled data [8, 13], this has several issues e.g. labels are highly biased by customers [17].

This paper presents a novel CSE method that uses large amounts of unlabeled data to mitigate the data limitation problem. The proposed method performs unsupervised representation learning based on self-training of input feature reconstruction as presented in natural language processing [18] and automatic speech recognition [19]. The main contribution of the proposed method is the introduction of a new loss function, called Multi-Format Prediction (MFP) loss, in the self-training step. The heuristic features widely used in CSE contain not only continuous features but also discrete ones like BoW of specific words, which takes 0 on most inputs. In this case, the conventional L1 loss used in self-training is not suitable since they easily lead to outputting the same 0 values on all inputs. MFP loss solves this issue by using two different loss functions to cover all features; the conventional L1 loss function and a weighted binary classification-based loss function are employed to handle continuous and discrete features with biases, respectively. The proposed method employs two-step training. First, an unsupervised pre-training model that reconstructs a masked part of the turn-level heuris-

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

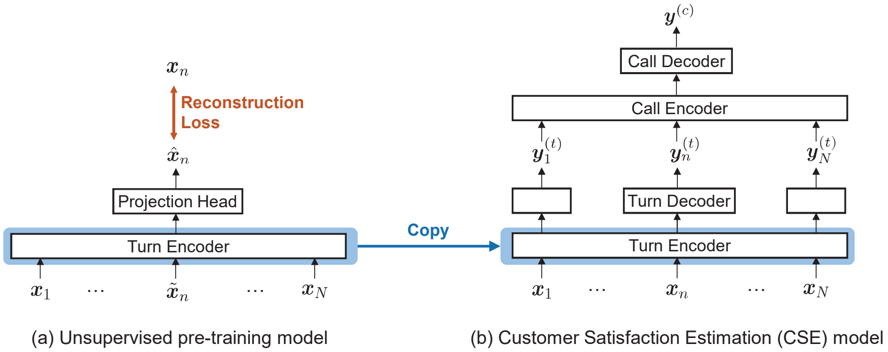

Fig. 1: Overview of the proposed unsupervised representation learning.

tic features is trained. The CSE model is then fine-tuned by the labeled training data using some of the parameters of the pre-training model as initial values. Experiments on real contact center calls reveal that the proposed method improves both call-level and turn-level CSE performance.

2. CUSTOMER SATISFACTION ESTIMATION USING A HIERARCHICAL MULTI-TASK MODEL

2.1. Model Architecture

This section describes the model and the input features of the conventional CSE based on the HMT model [16]. Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]$ be a sequence of turn-level features in a call, and N be the total number of customer turns. Turn-level and call-level CSE are formulated as estimating the series of turn-level labels $[l_1^{(t)}, \cdots, l_N^{(t)}]$ which correspond to the turn-level features, and call-level label $l^{(c)}$ from $\boldsymbol{X}$.

$$\hat{l}_n^{(t)} = \underset{l_n^{(t)}}{\arg\max}\, P(l_n^{(t)}|\boldsymbol{X}), \tag{1}$$

$$\hat{l}^{(c)} = \underset{l^{(c)}}{\arg\max}\, P(l^{(c)}|\boldsymbol{X}), \tag{2}$$

where $l_n^{(t)}, l^{(c)} \in L$; L is a set of customer satisfaction labels, e.g., $\{$*negative*, *neutral*, *positive*$\}$ in this paper.

The structure of the HMT model is shown in Fig. 1(b). The posterior probabilities of the turn-level label $\boldsymbol{y}_n^{(t)} = [P(l_n^{(t)} = \text{`}\textit{negative}\text{'}|\boldsymbol{X}), \cdots]$ are evaluated from a series of turn-level features using the turn encoder and the decoder. Those of the call-level labels, $\boldsymbol{y}^{(c)} = [P(l^{(c)} = \text{`}\textit{negative}\text{'}|\boldsymbol{X}), \cdots]$ are obtained from the output of the turn-level label estimation results.

All of the parts of the HMT model is jointly optimized by weighted sum of turn-level classification loss $\mathcal{L}^{(t)}$ and call-level classification loss $\mathcal{L}^{(c)}$,

$$\mathcal{L}_{\text{HMT}} = \alpha\mathcal{L}^{(t)} + (1-\alpha)\mathcal{L}^{(c)}, \tag{3}$$

where α is a HMT loss weight. $\mathcal{L}^{(t)}$ and $\mathcal{L}^{(c)}$ are evaluated by multi-class cross entropy.

Table 1: The 20 most class-specific words for each turn-level satisfaction label. The bold terms are used for BoW features.

negative	**because**, **don't/not**, **anything**, **why**, can, **right**, **wanted/want**, **nothing**, i've, **just**, do, **frustrating**, could, **issue**, understand, **what**, **shame**, **whenever**
positive	**thank/thanks**, **much**, **you**, **appreciate**, **very**, **perfect**, **help**, **so**, **awesome**, **ok**, **great**, bye, that's, your, **good**, be, those, all, should

2.2. Features

Turn-level feature $\boldsymbol{x}_n$ is composed of prosodic, lexical, and interactive features. Note that some features have been changed from [16] in this work because the target language is different.

Prosodic features they include statistics of log f_o, log power, and speech rate. The 18-dimensional features used are taken from [16] except for speech rate. Words-per-minute of the customer and the previous agent turn obtained by automatic speech recognition are used instead of mora-per-sec.

Interactive featues 11-dimensional interactive features that represent the length, rate, and frequency of the customer/agent turns and backchannels are used. They are the same as [16].

Lexical featues 29-dimensional word-based features are used as the lexical features; the total number of words in the current customer turn or previous agent turn, number of backchannel words (e.g. 'uh', 'hmm') in the current customer turn or previous agent turn, and 25-dimensional BoW of the target words in the current customer turn. The target words are selected as the class-specific words by the mutual information between the word and the turn-level customer satisfaction labels [20]. Several types of real contact center calls are used for the selection. The selected words shown in Table 1 and their inflections are used for BoW.

3. UNSUPERVISED REPRESENTATION LEARNING WITH MULTI-FORMAT PREDICTION LOSS

This section describes the proposed unsupervised representation learning with MFP loss which is inspired by Mockingjay [19]. The proposed method is overviewed in Fig. 1.

The pre-training model consists of a turn-level encoder and a projection head block. The hyper-parameters of the turn-level encoder are the same as those of the HMT model. The model estimates the reconstructed turn-level features $\hat{\boldsymbol{x}}_n$ of the masked region n from the input sequence including the masked features $\tilde{\boldsymbol{x}}_n$, as shown in Fig. 1(a). In the pre-training step, 15% of the input turns are randomly selected as the masked regions, and reconstruction loss between the original and the reconstructed features are evaluated simultaneously. Three types of masks, overwrite all to zero, replace with a random turn, and leave untouched, are applied in the proportions of 80%, 10%, and 10%, respectively. These enable the turn-level encoder to learn the distributions of the turn-level features, which makes it easier to capture the characteristic changes associated with turn-level and call-level customer satisfaction results.

The main difference to [19] is that the MFP loss is used instead of L1 loss for reconstruction loss. The distributions of the heuristic features are different for each feature, as illustrated in Fig. 2. For example, BoW-based lexical features follow a discrete distribution with most turns being zero. L1 loss is not suitable for reconstructing these biased distributions since they tend to lead to outputting the same 0 values on all inputs. MFP loss solves this problem by combining two different loss functions, as shown in Fig. 3. It evaluates reconstruction error via weighted Binary Cross Entropy (BCE) loss $\mathcal{L}_{\text{BCE}}$ for discrete distribution features and L1 loss $\mathcal{L}_{\text{L1}}$ for continuous distribution features,

$$\mathcal{L}_{\text{L1}} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i\in I_c} |x_{n,i} - \hat{x}_{n,i}|, \tag{4}$$

$$\begin{aligned}\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i\in I_d} \{ & w_{i,1} s(x_{n,i}) \log(\sigma(\hat{x}_{n,i})) \\ & + w_{i,0}(1 - s(x_{n,i})) \log(1 - \sigma(\hat{x}_{n,i}))\}\,, \end{aligned} \tag{5}$$

$$\mathcal{L}_{\text{MFP}} = \beta\mathcal{L}_{\text{L1}} + (1-\beta)\mathcal{L}_{\text{BCE}}, \tag{6}$$

where $\sigma(\cdot)$ is a sigmoid function and $s(\cdot)$ is a step function that is 1 if the input is more than 0, otherwise 0. $x_{n,i}, \hat{x}_{n,i}$ are the i-th dimension of the original and the reconstructed features $\boldsymbol{x}_n, \hat{\boldsymbol{x}}_n$, respectively. I_c, I_d are the indices sets of the feature dimensions that follow continuous and discrete distributions, respectively. $w_{i,0}$ and $w_{i,1}$ are the BCE weights on the i-th dimension which are given as the inverse values of frequencies of $s(x_{n,i})$ for all unlabeled data. β is the loss weight of MFP loss.

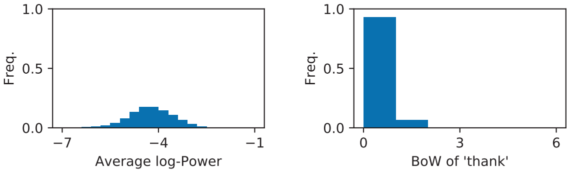

Fig. 2: Examples of the turn-level feature distributions.

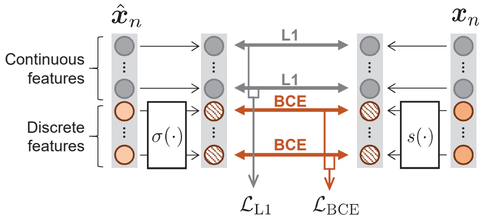

Fig. 3: Overview of MFP loss.

4. EXPERIMENTS

4.1. Dataset

The dataset was created from English calls in a retail customer services contact center. We randomly picked approximately 16k calls recorded from April to September 2020. Too-short/long calls which have less than 5 turns or more than 100 turns, respectively, were eliminated. Then hundreds of calls were randomly selected and annotated for the labeled dataset, while the rest were used as the unlabeled set. The number of turns and calls were 4466 and 170 in the labeled set; 388411 and 14782 in the unlabeled set, respectively.

Both call-level and turn-level customer satisfaction labels were annotated by three well-trained workers in the same way as in the previous work [16]. Fleiss' kappa values, measures for annotation agreements, were .82 for call-level and .51 for turn-level annotations, which suggested the annotations were sufficiently reliable. The ground truths were determined by harmonizing annotator-wise call-level and turn-level labels. Call-level ground truths were determined by the majority-voting of annotator-wise labels, and turn-level ones were *positive* or *negative* if at least one annotator gave *positive* or *negative* to that turn[1]. The distributions of call-level and turn-level ground truths in the labeled set are shown in Table 2.

4.2. Setups

Performance was evaluated by 5-fold cross-validation. Three folds were used as the training set, another one as the development, and the rest as the test set. In the pre-training step, 10% of the unlabeled dataset were used as the development set and

[1] We evaluated the majority-voting-based turn-labels as a preliminary experiments, which decreases call-level estimation performances with the HMT model.

Table 2: Class distributions in the labeled dataset.

	negative	*neutral*	*positive*
Turn	170	4096	200
Call	26	97	47

the rest 90% were the training set. The model parameters of the pre-trained model were the same for each evaluation in the cross-validation.

In turn-level feature extraction, frame length and frame shift of log f_o and log power were set at 64 ms and 5 ms, respectively. The dominant harmonic components method [21] was used for f_o extraction. A large-scale automatic speech recognition system which is trained by several hundreds of hours of transcribed real contact center calls was used to obtain words for the lexical features. All turn-level features were normalized against the unlabeled set.

The model structure and the training parameter of the CSE model were as follows. In the CSE model, the turn encoder consisted of 1-layer Fully-Connected (FC) and 1-layer bidirectional Long Short-Term Memory (LSTM) recurrent neural networks, both with 128 units. The turn decoder was 1-layer FC with 64 units. The call encoder was 1-layer FC and 1-layer unidirectional LSTM, both with 64 units. The call decoder was 1-layer FC with 16 units. The dropout rate was 0.2, and layer normalization was applied between each layer. The baseline was the flat-start training of the CSE model using just labeled data. The proposed methods applied unsupervised pre-training using the two loss functions, L1 and MFP loss, to the unlabeled dataset. The turn encoder of the pre-training model had the same structure as the CSE model and the prediction head was 1-layer FC with 128 units.

Minibatch size was 16 in pre-training and 3 in the flat-start or the fine-tuning steps. Adam was used for optimization, and the learning rates were 0.001 in pre-training and flat-start, and 0.0002 in fine-tuning. HMT loss weight α and MFP loss weight β were set to 0.8 and 0.9 as they yielded the highest performance from 0.1 to 0.9. Inverse class frequency was used as the class weight of turn-/call-level classification and binary classifications on MFP to mitigate data imbalance problems. Early stopping was triggered by the losses of the development set. The evaluation metrics were the overall accuracy and the macro-averaged F-measures of all classes.

4.3. Results and Discussions

Results are shown in Table 3. The pre-trained model with MFP loss showed better accuracy and macro F1 values than the baseline flat-start model in both tasks. On the other hand, the L1-based pre-trained model improved estimation performance in the turn-level task, but not in the call-level task. It is considered that our MFP loss proposal improved the reconstruction performance of various types of input features, which resulted in more stable fine-tuning.

Table 3: Estimation accuracies in turn-level and call-level customer satisfaction. Acc. and mF1 represent the overall accuracy and the macro-averaged F-measures of all classes, respectively.

		Turn		Call	
	P/T Loss	Acc.	mF1	Acc.	mF1
Flat-start	-	.857	.525	.571	.522
Pre-trained	L1	**.878**	**.546**	.571	.492
	MFP	.875	.543	**.647**	**.600**

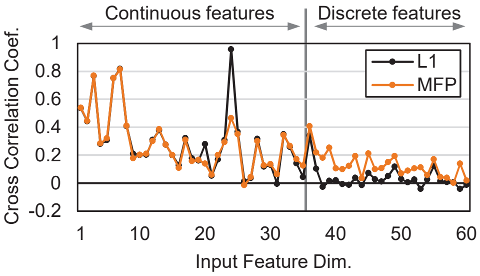

Fig. 4: Cross correlation coefficients between the original and the reconstructed features.

To enhance this discussion, feature reconstruction performances were compared to evaluate the cross-correlation coefficients between the original and the reconstructed features. Turn-by-turn reconstruction was used to get a series of reconstructed features; the reconstructed features in the n-th turn were yielded by the zero-masked n-th turn features and the original inputs in the rest turns. Results are shown in Fig. 4. The pre-trained model based on L1 loss presented moderate cross-correlation values in the reconstruction of the continuous features, while close to zero for most discrete features. However, the MFP loss-based model showed higher correlation values for the discrete features while keeping almost the same reconstruction performances as the L1-based model for continuous features. These indicate that the MFP loss makes it possible to acquire the characteristics of biased discrete features, which fails for the conventional L1 loss.

5. CONCLUSION

In this paper, we addressed turn-level and call-level CSE. To solve the issue of the lack of labeled data, a new CSE based on unsupervised representation learning was proposed. A new loss function, MFP loss, was also presented to improve the reconstruction performance of unsupervised training with various types of heuristic features. Experiments showed that the proposed scheme with MFP loss improved both call-level and turn-level estimation accuracy. Furthermore, the reconstruction results with our MFP loss demonstrated higher correlations with the original features.

6. REFERENCES

[1] Celestine C. Aguwa, Leslie Monplaisir, and Ozgu Turgut, "Voice of the customer: Customer satisfaction ratio based analysis," *Expert Systems with Applications*, vol. 39, no. 11, pp. 10112–10119, 2012.

[2] Geoffrey Zweig, Olivier Siohan, George Saon, Bhuvana Ramabhadran, Daniel Povey, Lidia Mangu, and Brian Kingsbury, "Automated quality monitoring for call centers using speech and NLP technologies," in *Proc. of NAACL*, 2006, pp. 292–295.

[3] Jordi Luque, Carlos Segura, Ariadna Sánchez, Martí Umbert, and Luis Angel Galindo, "The role of linguistic and prosodic cues on the prediction of self-reported satisfaction in contact centre phone calls," in *Proc. of INTERSPEECH*, 2017, pp. 2346–2350.

[4] Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu, "Mining call center conversations exhibiting similar affective states," in *Proc. of PACLIC*, 2016, pp. 545–553.

[5] Laurence Devillers, Christophe Vaudable, and Clément Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study.," in *Proc. of INTERSPEECH*, 2010, pp. 2350–2353.

[6] Jia Sun, Weiqun Xu, Yonghong Yan, Chaomin Wang, Zhijie Ren, Pengyu Cong, Huixin Wang, and Junlan Feng, "Information fusion in automatic user satisfaction analysis in call center," in *Proc. of IHMSC*, 2016, pp. 425–428.

[7] Narichika Nomoto, Masafumi Tamoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Kobashikawa, "Anger recognition in spoken dialog using linguistic and para-linguistic information," in *Proc. of INTERSPEECH*, 2011, pp. 1545–1548.

[8] Joseph Bockhorst, Shi Yu, Luisa Polania, and Glenn Fung, "Predicting self-reported customer satisfaction of interactions with a corporate call center," in *Proc. of ECML PKDD*, 2017, pp. 179–190.

[9] Youngja Park and Stephen C. Gates, "Towards real-time measurement of customer satisfaction using automatically generated call transcripts," in *Proc. of ACM*, 2009, pp. 1387–1396.

[10] Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kopparapu, "Event based emotion recognition for realistic non-acted speech," in *Proc. of TENCON*, 2015, pp. 1–5.

[11] Pengyu Cong, Chaomin Wang, Zhijie Ren, Huixin Wang, Yanmeng Wang, and Junian Feng, "Unsatisfied customer call detection with deep learning," in *Proc. of ISCSLP*, 2016, pp. 1–5.

[12] Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *Proc. of ICASSP*, 2019, pp. 5876–5880.

[13] Yelin Kim, Joshua Levy, and Yang Liu, "Speech sentiment and customer satisfaction estimation in socialbot conversations," in *Proc. of INTERSPEECH*, 2020, pp. 1833–1837.

[14] Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan, "Ordinal learning for emotion recognition in customer service calls," in *Proc. of ICASSP*, 2020, pp. 6494–6498.

[15] Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, and Yushi Aono, "Hierarchical LSTMs with joint learning for estimating customer satisfaction from contact center calls," in *Proc. of INTERSPEECH*, 2017, pp. 1716–1720.

[16] Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, Yushi Aono, and Tomoki Toda, "Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 715–728, 2020.

[17] Jeremy Auguste, Delphine Charlet, Geraldine Damnati, Frederic Bechet, and Benoit Favre, "Can we predict self-reported customer satisfaction from interactions?," in *Proc. of ICASSP*, 2019, pp. 7385–7389.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of NAACL-HLT*, 2019, pp. 4171–4186.

[19] Andy Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *Proc. of ICASSP*, 2020, pp. 6419–6423.

[20] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[21] Tomohiro Nakatani and Toshio Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3690–3700, 2004.