Time to Transfer: Predicting and Evaluating Machine-Human Chatting Handoff

Jiawei Liu,1 Zhe Gao,2 Yangyang Kang,2 Zhuoren Jiang,3 Guoxiu He,1 Changlong Sun,2,3 Xiaozhong Liu,4 Wei Lu1*

1Wuhan University, Wuhan, China
2Alibaba Group, China
3Zhejiang University, Hangzhou, China
4Indiana University Bloomington, Bloomington, United States

{iuaujames2017, guoxiu.he, weilu}@whu.edu.cn, {gaozhe.gz, yangyang.kangyy}@alibaba-inc.com, jiangzhuoren@zju.edu.cn, changlong.scl@taobao.com, liu237@indiana.edu

Abstract

Is chatbot able to completely replace the human agent? The short answer could be – “it depends...”. For some challenging cases, e.g., dialogue’s topical spectrum spreads beyond the training corpus coverage, the chatbot may malfunction and return unsatisfied utterances. This problem can be addressed by introducing the Machine-Human Chatting Handoff (MHCH) which enables human-algorithm collaboration.

To detect the normal/transferable utterances, we propose a Difficulty-Assisted Matching Inference (DAMI) network, utilizing difficulty-assisted encoding to enhance the representations of utterances. Moreover, a matching inference mechanism is introduced to capture the contextual matching features. A new evaluation metric, Golden Transfer within Tolerance (GT-T), is proposed to assess the performance by considering the tolerance property of the MHCH. To provide insights into the task and validate the proposed model, we collect two new datasets. Extensive experimental results are presented and contrasted against a series of baseline models to demonstrate the efficacy of our model on MHCH.

Introduction

Chatbot, as one of the recent successful applications in artificial intelligence, makes many people believe that it will be able to replace the human agent in the foreseeable future. Deep learning efforts from industry and academia boost the development of chatbot and have rewarded many inspiring accomplishments (High 2012; Zhou et al. 2020; Liu et al. 2020). However, Morar Consulting’s study suggests that only 21% respondents said they would like to be able to purchase goods and services from chatbots. Besides, due to the complexity of human conversation, automatic chatbots are not yet able to meet all the needs of users. The unsatisfied responses from algorithms may deteriorate user experience and cause business failure (Radziwill and Benton 2017). While scholars can consistently explore new models to further enhance the performance, in this study, we try to solve this problem from a perspective by combing chatbots with human agents. Specifically, we target the Machine-Human Chatting Handoff (MHCH), which aims to enable smooth and timely handoff from chatbot to human agent.

Figure 1 depicts an exemplar MHCH case of customer service for online sales in an E-commerce ecosystem. While chatbot successfully answers the first question from user (utter1 and utter2), unfortunately, it fails to address the second one decently (utter3 and utter4). In the MHCH task, the model should consistently monitor the progress of conversations, be able to predict the chatbot failures (potential algorithm failures, negative emotions from the user, etc.), and transfer the conversations to the human agent (a.k.a. handoff) to ensure the smooth dialogue. Recently, there are already several works focus on the human-machine cooperation for chatbots. For instance, Huang, Chang, and Bigham (2018) integrated crowds with multiple chatbots and voting systems. Rajendran, Ganhotra, and Polymenakos (2019) paid attention to transfer conversations to human agents once encountered new user behaviors. Different from these prior tasks, the MHCH task aims to sequentially label each utterance in the dialogue as normal or transferable based on the
dialogue content and its contextual information. The MHCH task can improve the ability of chatbot from two aspects. First, it can predict the potential risks of the dialogue and improve user satisfaction. Second, it helps to optimize the human agent resource allocation and reduce the business cost.

In this study, we propose a novel Difficulty-Assisted Matching Inference (DAMI) model for the MHCH task. The DAMI model is a hierarchical network that mainly consists of a difficulty-assisted encoding module and a matching inference mechanism. Considering the difficulty of the utterances affects the chance of handoff, we propose the difficulty-assisted encoding module. Besides the lexical and semantic information, it encapsulates lengths, part-of-speech tags, and term frequencies to learn and generate comprehensive representations of utterances. The matching inference module utilizes the contextual interaction to capture irrelevant or repeated utterances that may deteriorate user experience. The classification is based on the contextual representations and the integration of the utterance representations and matching features.

As for validating model performance for the MHCH task, none of the existing evaluation metrics seem to be very suitable because of the following reasons. **Imbalance**: The portion of normal utterances and transferable utterances is quite imbalanced. Conventional metrics, such as accuracy, will be problematic (Soda 2011). **Comprehensiveness**: Correctness of detecting the transferable utterances and coverage of transferable utterances are both important. Neither precision nor recall can meet this requirement. Moreover, the contribution of prediction correctness for the dialogues without handoff cannot be ignored. **Tolerance**: Although F score, AUC, and other similar metrics are designed for the comprehensive evaluation, they can be too rigid for this task. For instance, a slightly earlier handoff is acceptable and could not be considered as a complete mistake. Inspired by the Three-Sigma Rule (Pukelsheim 1994), we propose a novel evaluation metric **Golden Transfer within Tolerance (GT-T)** for the MHCH task, which takes the imbalance, comprehensiveness, and tolerance into consideration.

To summarise, our main contributions are three folds: (1) We propose a novel DAMI model utilizing difficulty-related and matching information for the MHCH task; (2) We propose a new evaluation metric GT-T to cope with the imbalance, comprehensiveness, and tolerance problem in the MHCH task; (3) To assist other scholars reproduce the experiment outcomes and further investigate this novel but important problem, two real-world customer service dialogue datasets are collected, labeled.

The experimental results demonstrate the superiority of our approach.

**Related Works**

In MHCH, a transferable or normal label is sequentially assigned to each utterance in a dialogue. To this end, the MHCH task can be defined as a supervised problem such as a classification problem or a sequence labeling problem. Classic classification algorithms, such as LSTM (Hochreiter and Schmidhuber 1997) and TextCNN (Kim 2014), learn the representation of each utterance and predict the label straightforwardly. Recently, pre-trained language representations, such as BERT (Devlin et al. 2019), have improved many NLP downstream tasks by providing rich semantic features. However, for instance, in a customer service scenario, customer’s satisfaction can change gently with the dialogue moving forward. A decent model for the MHCH task should be able to sense user’s emotional change or dissatisfaction in the earliest step, and a delayed handoff may cause user’s impatience and abandonment. Compared to classic text classification methods, Yang et al. (2016), Chen et al. (2018), Raheja et al. (2019), and Dai et al. (2020) proposed different attention mechanisms to capture contextual information or dependencies among labels. Besides attention mechanisms, Yu et al. (2019) presented an adapted convolutional recurrent neural network that models the interactions between utterances of long-range context. Majumber et al. (2019) proposed a method based on recurrent neural networks that keeps track of the individual party states throughout the conversation and uses this information for emotion classification.

The mismatched or even irrelevant answers in some cases deteriorate the customer experience (Radziwill and Benton 2017). How to capture the matched or mismatched information have been widely studied in response selection of IR-based chatbots or sentiment analysis (Chen et al. 2017; Liu et al. 2018; Shen et al. 2018; Mao et al. 2019). Besides users’ emotions, text difficulty or readability affects the chance of handoff. There are a large number of researches based on machine learning methods to evaluate text difficulty (Collins-Thompson and Callan 2004; Hancke, Vajjala, and Meurers 2012; Jiang et al. 2018). Generally speaking, it is mainly related to the word (frequency, length, etc.), grammar, syntax, sentence length, etc. (Nelson et al. 2012; Benjamin 2012). Liao, Srivastava, and Kapanipathi (2017) proposed a data-driven method to calculate the dialogue complexity and optimize the dispatching of human agents.

Since fully-automated chatbots still have limitations in the foreseeable future, a promising direction is to combine chatbots with human agents. Earlier research, the crowd-powered conversational assist architecture, Evorus (Huang, Chang, and Bigham 2018), integrated crowds with multiple chatbots and voting systems. Rajendran, Ganhotra, and Polymenakos (2019) paid attention to transfer dialogues to human agents once encountered new user behaviors. After a period of deployment, they could automate itself with decent conversation quality. Different from them, we mainly focus on detecting transferable utterances which are one of the keys to improve the user satisfaction and optimize the human agent resource allocation.

**Methodology**

**Our Model**

Formally, we assume that a dialogue $D$ is composed of a sequence of utterances $[u_1, \ldots, u_t, \ldots, u_L]$ with corresponding handoff labels $[y_1, \ldots, y_t, \ldots, y_L]$, where $y_t \in \mathcal{Y}$. $\mathcal{Y}$ is the set of handoff labels (normal and transferable). Transferable indicates the dialogue should be transferred to the human
agent, whereas normal indicates there is no need to transfer. As shown in Figure 2, our model mainly consists of three components: Difficulty-Assisted Encoding, Matching Inference Mechanism, and Utterance Labeling Layer.

**Difficulty-Assisted Encoding**: The difficulty of the utterance affects the chance of handoff. On the one hand, the chatbot may have difficulty to understand and answer the questions raised by customers. On the other hand, customers may have difficulty to understand or capture the useful information of the compound or lengthy responses given by the chatbot. Remarkably, the difficulty or readability of the text is related to the word (frequency, length, etc.), grammar, syntax, sentence length, etc. (Benjamin 2012). Thus, besides the lexical and semantic information, we make use of positions, term frequencies, and part-of-speech tags to enhance the representation of utterance difficulty.

Suppose $u_t = [w_1, ..., w_{|u_t|}]$ represents a sequence of words in $u_t$. These words are mapped into corresponding word embeddings $E_u^{w} \in \mathbb{R}^{d \times |u_t|}$, where $d$ is the word embedding dimension. Positional embedding (Vaswani et al. 2017), which can encode order dependencies, is utilized to represent the increasing length information. $E_p^{u} \in \mathbb{R}^d$ is the positional embedding of $w_1$ and $E_p^{u} \in \mathbb{R}^{d\times |u_t|}$ is the positional embeddings list of words in $u_t$. The one-hot representations of part-of-speech tags $E_s^{u} \in \mathbb{R}^{n \times |u_t|}$ of words in $u_t$ are also generated. Note that $n$ is the number of part-of-speech tag categories. Hence, the overall embedding $E_{u_t} \in \mathbb{R}^{(2d + n) \times |u_t|}$ of $u_t$ is:

$$E_{u_t} = E_u^{w} \oplus E_p^{u} \oplus E_s^{u}$$

where $\oplus$ denotes concatenation operation. We utilize a bidirectional LSTM (BiLSTM) for obtaining the output of every time step $a_t \in \mathbb{R}^{2k \times |u_t|}$ and the last state concatenation of forward and backward $s_t \in \mathbb{R}^{2k}$, where $k$ is the number of hidden units of LSTM cell. Briefly, we have:

$$o_t, s_t = \text{BiLSTM}(E_{u_t})$$

In sales customer service dialogue, the roles of different participants would exhibit different characteristics (Song et al. 2019). Customers tend to use more succinct phrases, whereas agents tend to use more elaborated sentences (Liao, Srivastava, and Kapanipathi 2017). Thus, we use different vectors to capture different roles’ difficulty features:

$$W^c_{u} = r_1W^c_{\omega} + (1 - r_1)W^u$$
$$b^c_{u} = r_2b^c_{\omega} + (1 - r_2)b^c_{u}$$
$$g^c_{u} = r_3g^c_{\omega} + (1 - r_3)g^c_{u}$$

where $W^{(a,c)} \in \mathbb{R}^{z \times d}$, $b^{(a,c)} \in \mathbb{R}^{z}$ and $g^{(a,c)} \in \mathbb{R}^{z \times |u_t|}$ are trainable parameters shared across utterances of agents and customers, respectively. $z$ is the attention size. Since there are two roles in the MHCH task, we use $r_1 \in \{0, 1\}$ to represent the role of utterance $u_t$. Intuitively, the word frequency can also reflect the text difficulty: rarely used words (low word frequency) usually increase the understanding difficulty (Chen and Meurers 2018). Thus, we employ a term-frequency-adjusted attention mechanism to further enhance the difficulty-assisted representation $a_t \in \mathbb{R}^{2k}$:

$$g_{u_t} = \text{tanh}(W^c_{\omega}E_{u_t} + b^c_{u})$$
$$\alpha_{u_t} = \text{softmax}((1 - tf_{u_t}) \cdot g^r_{u_t} \cdot g^r_{\omega})$$
$$a_t = \sum_{t=1}^{\frac{|u_t|}{n}} o_t \cdot \alpha_{u_t}$$

where $g_{u_t} \in \mathbb{R}^{z \times |u_t|}$ is the hidden representations and encoded by a fully connected layer. Then we compute the normalized difficulty weights $\alpha_{u_t}$ with a local vector $g^r_{u_t}$ and normalized term frequencies $tf_{u_t} \in [0, 1]$ of utterance $u_t$.

Meanwhile, the customer’s emotion is an indicator of service quality, which is potentially helpful for the handoff judgment. Hence, we also calculate an emotional polarity score $e_t \in \mathbb{R}$ for utterance $u_t$ by SnowNLP.\footnote{https://github.com/isnowfy/snownlp}

Figure 2: Overall architecture of Difficulty-Assisted Matching Inference (DAMI) network.
Finally, we get the utterance representation vector $v_t \in \mathbb{R}^k$ ($K = 4k + 1$) of $u_t$ as follow:
\begin{equation}
  v_t = s_t \oplus a_t \oplus e_t
\end{equation}

**Matching Inference Mechanism:** The customer tends to turn to a human agent for help when the chatbot gives an unsatisfactory or irrelevant response. Practically, it is very difficult and paradoxical to directly judge whether the given answer is satisfactory. Nevertheless, the repeated utterances (the same responses or semantically similar questions) would increase customer’s dissatisfaction. Based on this observation, we make the present utterance interact with preceding utterances to get the contextual matching features. For utterance $u_t$, we calculate its matching features $m_t$ by:
\begin{equation}
  m_t = v_t^\top [v_1, v_2, ..., v_{t-1}]
\end{equation}

By masking out the future information of current utterance, the matching features of dialogue $D$ can be represented as $V_M = [m_1, ..., m_{t-1}, ..., m_L]$, where $V_M \in \mathbb{R}^{L\times L}$ is a lower triangular matrix with the diagonal value removed. Then we use a fully connected layer to further integrate the matching features with utterance vector $\hat{v}_t \in \mathbb{R}^k$:
\begin{equation}
  \hat{v}_t = \text{ReLU}(W_\gamma(m_t \oplus v_t) + b_\gamma)
\end{equation}
where $\gamma \in \mathbb{R}^{k \times (K+T)}$ and $b_\gamma \in \mathbb{R}^k$. ReLU denotes the rectified linear unit activation ($\text{ReLU}(x) = \max(0, x)$).

**Utterances Labeling:** The tendency of handoff also depends on the dialogue context. Thus, to connect the information flow in the dialogue, we feed the integrated representations into an LSTM:
\begin{equation}
  h_t = \text{LSTM}(\hat{v}_t, h_{t-1}, c_{t-1})
\end{equation}
where $h_t \in \mathbb{R}^k$ is the hidden state of LSTM for the utterance $u_t$ and $c_{t-1}$ is the memory cell state at time-step $t-1$.

Furthermore, to estimate a long-range dependency, we employ an attention mechanism [Majumder et al. 2019] to compute the context representation $\hat{c}_t \in \mathbb{R}^k$ of utterance $u_t$:
\begin{equation}
  \alpha_t = \text{softmax}(h_t^\top W_\alpha [h_1, h_2, ..., h_{t-1}])
\end{equation}
\begin{equation}
  \hat{c}_t = \alpha_t [h_1, h_2, ..., h_{t-1}]^\top
\end{equation}
where $[h_1, h_2, ..., h_{t-1}]$ are the hidden states for the preceding utterances of utterance $u_t$ and $W_\alpha \in \mathbb{R}^{k \times k}$. Then we utilize a fully connected layer to get the context-aware representation $h_t \in \mathbb{R}^k$:
\begin{equation}
  h_t = \text{ReLU}(W_\gamma(h_t \oplus \hat{c}_t) + b_\gamma)
\end{equation}
where $W_\gamma \in \mathbb{R}^{k \times 2k}$ and $b_\gamma \in \mathbb{R}^k$.

Since there are no dependencies among labels, we simply use a softmax classifier for performing handoff prediction:
\begin{equation}
  p_t = \text{softmax}(W_\gamma h_t + b_\gamma)
\end{equation}
where $W_\gamma \in \mathbb{R}^{2 \times k}$ and $b_\gamma \in \mathbb{R}^2$ is the predicted handoff probability distribution of $u_t$.

**End-to-End Training:** We use categorical cross-entropy as the measure of loss ($\mathcal{L}$) during training:
\begin{equation}
  \mathcal{L} = -\frac{1}{T} \sum_{i=1}^{I} \sum_{t=1}^{T} \log p_{i,t}(y_{i,t}|D_i, \Theta) + \frac{\delta}{2} \|\Theta\|_2^2
\end{equation}

where $I$ is the number of dialogues, $D_i$ is the $i$-th dialogue, $p_{i,t}$ is the probability distribution of handoff label of $t$-th utterance of $D_i$, $y_{i,t}$ is the expected class label of $t$-th utterance of $D_i$, $\delta$ denotes the $L_2$ regularization weight and $\Theta$ denotes all the trainable parameters of model. We use backpropagation to compute the gradients of the parameters, and update them with Adam (Kingma and Ba 2014) optimizer.

**Golden Transfer within Tolerance**

For the MHCH tasks, it is vital to detect all of the transferable utterances. Meanwhile, predicting normal utterances as transferable will also cause a waste of human resources. In order to comprehensively measure the model performance for predicting the transferable or normal label of an utterance, we adopt F1, Macro F1, and AUC. However, the evaluation metrics like F1, Macro F1, and AUC have several limitations. First, if the predicted transferable utterance is not the true transferable utterance, these metrics will treat it as an error, no matter how close these two utterances are. This characteristic can be too rigid for the MHCH task. Second, F1 and AUC have no reward for the correct prediction of the dialogues without handoff.

To address above problems, we propose a new evaluation metric, namely **Golden Transfer within Tolerance (GT-T).**

Inspired by the Three-Sigma Rule [Pukelsheim 1994], the proposed GT-T allows a “biased” prediction within the tolerance range $T \in \mathbb{N}$. Specifically, in this paper, we set $T$ to range from 1 to 3 corresponding to GT-I, GT-II, and GT-III.

Suppose a dialogue session contains $L \in \mathbb{N}$ utterances and $N \in \mathbb{N}$ true transferable utterances, where $N \leq L$. The number of predicted transferable utterances is $M \in \mathbb{N}$, where $M \leq L$. The set of true transferable utterances’ positions is $Q = \{q_1, q_2, ..., q_N\}$, where $0 \leq q_j \leq L - 1$ and $j \in \{1, 2, ..., N\}$. The set of predicted transferable utterances’ positions is $P = \{p_1, p_2, ..., p_M\}$, where $0 \leq p_i \leq L - 1$ and $i \in \{1, 2, ..., M\}$. We compute this session’s GT-T score as:
\begin{equation}
  \text{GT-T}(T) = \frac{1}{M} \sum_{i=1}^{M} \max_{1 \leq j \leq N} \exp\left(\frac{1}{\lambda \cdot \text{sgn}((\Delta_{ij} - 1) \cdot \frac{\Delta_{ij}^2}{2(T+1)^2}) \cdot \delta^2\right), \text{others})
\end{equation}

Note that $\epsilon$ is an infinitesimal, $\text{sgn}$ represents the signum function and $\Delta_{ij} = p_i - q_j, \lambda \in (-1, 1)$ is an adjustment coefficient for the early handoff and the delayed handoff. An early handoff may improve the user experience but exacerbate the human agent resource scarcity. In turn, a delayed handoff may mitigate the human agent resource scarcity but sacrifice the user experience. $\lambda > 0$ means that the user experience is more important than the human agent resource. $\lambda < 0$ means that the human agent resource is more important than the user experience.

There are three advantages of the proposed GT-T metric:

1. **Focusing on handoff.** Since the transferable utterances are a minority in the dialogue with handoff (refer to Table 1), conventional metrics, such as accuracy, will be problematic as the **imbalance** of classes. Contrarily, GT-T mainly focuses on transferable. Thus it will not be affected by normal
utterances. (2) **Comprehensiveness.** For the MHCH task, the correctness of detecting the transferable utterances and coverage of all transferable utterances are both important. Moreover, handoff does not necessarily happen in all cases, which means the correct judgment of the dialogues without handoff should be rewarded. (3) **Tolerance.** Although F score, AUC, and other similar metrics are designed to comprehensively measure the model, they can be too rigid for the MHCH task. For example, if there is a dialogue with true labels sequences \( Q_0 = [0, 0, 0, 0, 0, 1] \), it means that the dialogue needs to be transferred to the human agent at the sixth round. The F1, Macro-F1, AUC, GT-I, GT-II, and GT-III all scored 1 for prediction \( P_I = [0, 0, 0, 0, 0, 1] \). If there are enough human agents resources or the customer’s patience is poor, the prediction, \( P_2 = [0, 0, 0, 0, 1, 0] \), can be acceptable and shouldn’t be considered as a complete mistake. In such situation, GT-I, GT-II, and GT-III scores (\( \lambda = 0 \)) of \( P_2 \) are 0.61, 0.88, and 0.95, while the F1, Macro-F1, and AUC scores of \( P_2 \) are 0, 0.4, and 0.4. Obviously, GT-T is more suitable for MHCH task.

**Datasets**

To the best of our knowledge, there is no publicly available dataset for the MHCH task. To address this problem, we propose two Chinese sales customer service dialogue datasets, namely **Clothing** and **Makeup**, which are collected from Taobao⁴, one of the largest decentralized E-commerce platforms in the world. Clothing is a corpus with 3,500 dialogues in the clothing domain and Makeup is a corpus with 4,000 dialogues in the makeup domain. User information is removed from both datasets. Each dialogue was annotated by an expert, and at least half of the dialogues were checked randomly by another expert. Five experts participated in the annotation work. Every utterance of the dialogue was assigned a transferable or normal label precisely without considering early or delayed handoff. Since the human agent is not involved in the dialogue, there may be multiple transferable utterances in each dialogue. Besides the explicit demand for human agent from the customer, an utterance will also be annotated as transferable if the customer receives an unsatisfactory answer or expresses negative emotions. When the customer or chatbot repeats semantically similar utterances, the repeated utterance may also be annotated as transferable.

A summary of statistics, including Kappa value (Snow et al. 2008) and inner annotator agreement measured by F-score for both datasets, are shown in Table 1. From the Table, we can see that the datasets are highly imbalanced in terms of label distributions and the normal utterances occupy the largest proportion. Meanwhile, the length of the utterances content varies greatly. What’s more, the distributions of transferable utterances’ relative locations in the dialogue between the two datasets are significantly different \( p < 0.001 \) by two-tailed Mann-Whitney U test. From the above observations, it is clear that these two datasets are very challenging for the MHCH task.

| Statistics items | Clothing | Makeup |
|------------------|----------|--------|
| # Dialogues      | 3,500    | 4,000  |
| # Normal dialogues| 274      | 255    |
| # Dialogues with handoff | 3,226    | 3,745  |
| # Utterances     | 35,614   | 39,934 |
| # Normal utterances | 28,901   | 32,488 |
| # Transferable utterances | 6,713    | 7,446  |
| Avg# Utterances per dialogue | 10.18    | 9.98   |
| Avg# Tokens per utterance | 17.71    | 21.45  |
| Std# Tokens of utterances  | 5.31     | 5.29   |

Table 1: Statistics of the datasets we collected

### Experiments and Results

#### Experimental Settings

All of the tokenization and part-of-speech tagging are performed by a popular Chinese word segmentation utility called jieba⁵. After preprocessing, the datasets are partitioned for training, validation and test with an 80/10/10 split. For all the methods, we apply fine-tuning for the word vectors. The word vectors are initially trained on Clothing and Makeup corpora by using CBOW (Mikolov et al. 2013). The dimension of word embedding is set as 200 and the vocabulary size is 19.5K. Other trainable model parameters are initialized by sampling values from the Glorot uniform initializer (Glorot and Bengio 2010). Hyper-parameters of our model and baselines are tuned on the validation set. The sizes of LSTM hidden state \( k \), attention units \( z \) are all 128, and batch size is set as 128. The dropout rate (Srivastava et al. 2014) is 0.75 and the number of epochs is set as 30. The \( L_2 \) regularization weight is \( 10^{-4} \). Finally, we train the models with an initial learning rate of 0.0075. All the methods are implemented by Tensorflow⁶ and run on a server configured with a Tesla V100 GPU, 8 CPU, and 16G memory.

#### Experimental Results

We compare our proposed model with the following state-of-the-art dialogue classification models, which mainly come from similar tasks, such as dialogue sentiment classification and dialogue act classification. We briefly describe these baseline models below: 1) **HRN** (Lin et al. 2015): It uses a bidirectional LSTM to encode utterances, which are then fed into a standard LSTM for context representation. 2) **HAN** (Yang et al. 2016): HAN is a hierarchical network, which has two levels of attention mechanisms, applied at the words and utterances. 3) **BERT** (Devlin et al. 2019): It uses a pre-trained BERT model to construct the single utterance representations for classification. 4) **CRF-ASN** (Chen et al. 2018): It extends the structured attention network to the linear-chain conditional random field layer, which takes both contextual utterances and corresponding dialogue acts into

⁴https://www.taobao.com/
⁵https://pypi.org/project/jieba/
⁶https://www.tensorflow.org/
Table 2: Performance (%) on Clothing and Makeup datasets. ** indicates statistical significance at p < 0.01 level compared to the best performance of baselines. λ of GT-T is 0.

| Models         | Clothing | | Makeup | |                  |
|----------------|----------|--------|--------|--------|-----------------|
|                | F1       | Macro | F1     | AUC    | GT-I | GT-II | GT-III | F1     | Macro | F1     | AUC    | GT-I | GT-II | GT-III |
| HRN            | 57.29    | 73.45 | 86.53  | 62.33  | 71.75 | 76.46 | 58.21  | 74.10  | 87.16 | 62.33  | 72.51 | 78.06 |
| HAN            | 58.17    | 74.12 | 86.76  | 62.93  | 72.17 | 76.72 | 60.08  | 75.29  | 87.21 | 65.44  | 74.89 | 78.87 |
| BERT           | 56.03    | 72.93 | 83.78  | 59.27  | 68.07 | 73.10 | 56.98  | 73.25  | 81.65 | 61.45  | 71.01 | 76.48 |
| CRF-ASN        | 57.62    | 73.35 | 84.62  | 61.45  | 72.58 | 77.95 | 56.81  | 73.59  | 83.87 | 63.65  | 74.24 | 79.84 |
| HBLSTM-CRF     | 59.02    | 74.39 | 86.46  | 63.61  | 73.72 | 78.82 | 60.11  | 75.43  | 86.08 | 67.01  | 76.29 | 81.16 |
| DialogueRNN    | 59.04    | 74.34 | 86.94  | 63.10  | 73.75 | 79.02 | 61.33  | 76.07  | 88.45 | 66.34  | 75.98 | 81.15 |
| CASA           | 59.73    | 74.74 | 86.99  | 64.84  | 74.89 | 79.66 | 60.38  | 75.73  | 87.98 | 67.79  | 76.96 | 81.79 |
| LSTM + LCA     | 61.81    | 76.09 | 85.79  | 66.38  | 76.27 | 81.07 | 62.06  | 76.61  | 88.93 | 67.34  | 76.87 | 81.72 |
| CESta          | 60.47    | 75.15 | 86.22  | 63.97  | 74.63 | 79.63 | 60.24  | 75.22  | 87.14 | 65.16  | 75.91 | 81.49 |

DAMi (Our Model) 67.19** 79.44** 91.23** 72.83** 81.27** 85.51** 69.26** 80.90** 92.31** 71.80** 79.60** 84.05**

** Table 2: Performance (%) on Clothing and Makeup datasets. ** indicates statistical significance at p < 0.01 level compared to the best performance of baselines. λ of GT-T is 0.

As shown in Table 2, our approach achieves the best performance against the strong baselines on the Clothing and Makeup datasets for the MHCH task. We also conduct Wilcoxon signed-rank tests with continuity correction between our method with the baselines (15 runs), and the results show the improvements are significant with p < 0.01. These experimental results demonstrate the effectiveness of utilizing matching and difficulty information to enhance the hierarchical network capability for the MHCH task. Meanwhile, we also find that, the performance of baseline models will change in different magnitudes with the change of λ. Taking GT-II on Makeup dataset as an example, LSTM + LCA performs better than CASA when λ ≤ −0.5, while it is opposite when λ ≥ −0.25. Overall, our models perform relatively better than baseline models.

Analysis and Discussion

Ablation study. To evaluate the contributions of the proposed modules in our model, we implement several model variants for ablation tests respectively. - **Emotion** gets the utterance representation without concatenating the emotional score. - **Matching** removes the Matching Inference mechanism. - **Difficulty-Assisted** replaces the Difficulty-Assisted Encoding with one layer BiLSTM. + **Attention** replaces the Difficulty-Assisted Encoding with one layer BiLSTM enhanced by self-attention. The results are recorded in the last four rows of Table 2. For the sake of fairness, we keep the other components in the same settings when modifying one module.

We can observe that - **Emotion** performs well but still worse than DAMI, indicating that concatenating emotional information is helpful but not noticeable compared with other components. Without **Matching** or **Difficulty-Assisted**, the performances drop a lot but are still better than baselines under a few indicators, which proves the effectiveness of these two modules, respectively.

Since our Difficulty-Assisted module is attention-based, we replace it with an attention mechanism. Working with Matching Mechanism, it also outperforms many baselines but not as competent as the full version of DAMI. By removing **Matching** and **Difficulty-Assisted** simultaneously, we can observe that...
Dialogue Content | LSTM + LCA | DAMI
---|---|---
C1: ***URL of a product***  | Normal | T
C2: Buy two get one free. | Normal | T
A1: Dear, you can add your favorite three goods to shopping cart directly. The system will cut the price automatically when you pay! | Normal | T
C3: Can I buy three of these goods at once? | Normal | T
A2: Dear, there will be a specific introduction on the detail page. If we have buy-two-get-one-free promotions, the system will cut the price automatically when you buy three goods. | Normal | T
C4: How to enjoy “buy one get one free”? | Normal | T
A3: Dear, you can pay attention to the front page of out store. […] A long text introducing a series of sales activities and preferential rules … | Transferable | T
(A few moments later.)
A3: Dear, your order has not been paid yet. We will give priority to ship your goods if you pay early. | Normal | T
C5: I don’t understand how to place orders. | Transferable | T
A4: Beauties, the detail page will have instructions. If the problem has not been resolved, please type “manual”. | Normal | T
C6: ***URL of a product***  | Normal | T
C7: Is any human here? | Transferable | T
(…Human agent intervenes this dialogue…)
Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work is supported by the National Natural Science Foundation of China (71673211), the National Key R&D Program of China (2020YFC0832505), and Alibaba Group through Alibaba Research Intern Program and Alibaba Research Fellowship Program.

References

Benjamin, R. G. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. Educational Psychology Review 24(1): 63–88.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced LSTM for Natural Language Inference. In Proceedings of ACL-2017, 1657–1668.

Chen, X.; and Meurers, D. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. Journal of Research in Reading 41(3): 486–510.

Chen, Z.; Yang, R.; Zhao, Z.; Cai, D.; and He, X. 2018. Dialogue act recognition via crf-attentive structured network. In Proceedings of SIGIR-2018, 225–234. ACM.

Collins-Thompson, K.; and Callan, J. P. 2004. A language modeling approach to predicting reading difficulty. In Proceedings of NAACL-2004, 193–200.

Dai, Z.; Fu, J.; Zhu, Q.; Cui, H.; Li, X.; and Qi, Y. 2020. Local Contextual Attention with Hierarchical Structure for Dialogue Act Recognition. arXiv preprint arXiv:2003.06044.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-2019, 4171–4186.

Glorot, X.; and Bengio, Y. 2010. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of AISTATS-2010, 249–256.

Hancke, J.; Vajjala, S.; and Meurers, D. 2012. Readability classification for German using lexical, syntactic, and morphological features. In Proceedings of COLING-2012, 1063–1080.

High, R. 2012. The era of cognitive systems: An inside look at IBM Watson and how it works. IBM Corporation, Redbooks.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. Neural computation 9(8): 1735–1780.

Huang, T.-H.; Chang, J. C.; and Bigham, J. P. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In Proceedings of CHI-2018, 1–13.

Jiang, Z.; Gu, Q.; Yin, Y.; and Chen, D. 2018. Enriching word embeddings with domain knowledge for readability assessment. In Proceedings of COLING-2018, 366–378.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In Proceedings of EMNLP-2014, 1746–1751.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kumar, H.; Agarwal, A.; Dasgupta, R.; and Joshi, S. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In Proceedings of AAAI-2018.

Liao, Q. V.; Srivastava, B.; and Kapanipati, P. 2017. A Measure for Dialog Complexity and its Application in Streamlining Service Operations. arXiv preprint arXiv:1708.04134.

Lin, R.; Liu, S.; Yang, M.; Li, M.; Zhou, M.; and Li, S. 2015. Hierarchical recurrent neural network for document modeling. In Proceedings of EMNLP-2015, 899–907.

Liu, C.; Jiang, S.; Yu, H.; and Yu, D. 2018. Multi-turn inference matching network for natural language inference. In Proceedings of NLPCC-2018, 131–143.

Liu, Q.; Chen, Y.; Chen, B.; Lou, J.-G.; Chen, Z.; Zhou, B.; and Zhang, D. 2020. You impress me: Dialogue generation via mutual persona perception. Proceedings of ACL-2020.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. Journal of machine learning research 9(Nov): 2579–2605.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In Proceedings of AAAI-2019, volume 33, 6818–6825.

Mao, G.; Su, J.; Yu, S.; and Luo, D. 2019. Multi-Turn Response Selection for Chatbots With Hierarchical Aggregation Network of Multi-Representation. IEEE Access 7: 111736–111745.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, 807–814.

Nelson, J.; Perfetti, C.; Liben, D.; and Liben, M. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Council of Chief State School Officers, Washington, DC.

Pukelsheim, F. 1994. The three sigma rule. The American Statistician 48(2): 88–91.

Radziwill, N.; and Benton, M. 2017. Evaluating Quality of Chatbots and Intelligent Conversational Agents. Software Quality Professional 19(3): 25.

Rahela, V.; and Tetreault, J. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In Proceedings of NAACL-2019, 3727–3733.

Rajendran, J.; Ganhotra, J.; and Polymenakos, L. C. 2019. Learning End-to-End Goal-Oriented Dialog with Maximal User Task Success and Minimal Human Agent Use. Transactions of the Association for Computational Linguistics 7: 375–386.
Shen, C.; Sun, C.; Wang, J.; Kang, Y.; Li, S.; Liu, X.; Si, L.; Zhang, M.; and Zhou, G. 2018. Sentiment classification towards question-answering with hierarchical matching network. In Proceedings of EMNLP-2018, 3654–3663.

Snow, R.; O’connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In Proceedings of EMNLP-2008, 254–263.

Soda, P. 2011. A multi-objective optimisation approach for class imbalance learning. Pattern Recognition 44(8): 1801–1810.

Song, K.; Bing, L.; Gao, W.; Lin, J.; Zhao, L.; Wang, J.; Sun, C.; Liu, X.; and Zhang, Q. 2019. Using Customer Service Dialogues for Satisfaction Analysis with Context-Assisted Multiple Instance Learning. In Proceedings of EMNLP-2019, 198–207.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15(1): 1929–1958.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In Proceedings of NeurIPS-2017, 5998–6008.

Wang, Y.; Zhang, J.; Ma, J.; Wang, S.; and Xiao, J. 2020. Contextualized Emotion Recognition in Conversation as Sequence Tagging. In Proceedings of SIGdial-2020, 186–195.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In Proceedings of NAACL-2016, 1480–1489.

Yu, Y.; Peng, S.; and Yang, G. H. 2019. Modeling Long-Range Context for Concurrent Dialogue Acts Recognition. In Proceedings of CIKM-2019, 2277–2280. ACM.

Zhou, L.; Gao, J.; Li, D.; and Shum, H.-Y. 2020. The design and implementation of xiaoice, an empathetic social chatbot. Computational Linguistics 46(1): 53–93.