Interspeech 2022
18-22 September 2022, Incheon, Korea

Emotion-Shift Aware CRF for Decoding Emotion Sequence in Conversation

Chun-Yu Chen, Yun-Shao Lin, Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Taiwan

adam@gapp.nthu.edu.tw, yunshaolin@gmail.com, cclee@ee.nthu.edu.tw

Abstract

Emotion recognition in conversation (ERC) is an increasingly important topic as it improves user experiences when adopting speech technology in our daily life. In this work, we propose an emotion-shift aware decoder based on formulation of conditional random field (CRF) to address the perennial issue of poor performances when handling emotion shift in dialogues. We conduct speech emotion recognition experiments on the IEMOCAP and the NNIME and achieve a 74.47% unweighted accuracy, which is the current state-of-the-art performance in the four class emotion recognition on the IEMOCAP. This is also the first work for ERC on the NNIME that obtains an outstanding performance of 61.02% weighted accuracy.

Index Terms: speech emotion recognition, conversation, conditional random field, emotion shift

1. Introduction

Research on emotion recognition in conversation (ERC) [1, 2] is becoming important due to the increasing proliferation and usage of speech technology in our daily life, where conversation is the dominant form of our everyday interactions [3]. Unlike models of isolated per-utterance emotion recognition, models of ERC require consideration of contextual history to better predict the emotion of a current utterance. A prevalent direction of research in ERC is to model emotion transitions of the dyad in a conversation. Interlocutor's emotion is influenced by one other over time and by one's own past history [4]. These contextual dependencies result in two forms of emotion transition in dialog: the natural persistence of an emotion (*inertia*) [5] and the eventual switch out of such a sustained state (*shift*) [6].

Many works have showed improved recognition rates when modeling emotion transitions for ERC [7, 8, 9, 10, 11]. These methods are broadly categorized into two major types of approaches. The first involves modeling utterance-to-emotion and conversation context simultaneously in a single sophisticated architecture. Notable works includes: Hazarika et al. used connected memory networks (ICON) to model the relation between the current speaker and the other speaker for improved recognition [7]; Yeh et al. designed an attention mechanism that converts contextual information into the learned vocal representation (IAAN) [8]; Poria et al. proposed a hierarchical RNN framework (DialogueRNN) that keeps track of the individual speaker states throughout the conversation to perform emotion classification [9]; Shen et al. combined the graph-based neural networks and recurrence-based neural networks (DAG-ERC) to model the information flow between long-distance conversation background and nearby context [10].

In contrast to the above works, which use complicated neural network architectures and consider mostly short-term local context between interlocutors in a conversation, a more recent line of research is to decouple this into two separate components: a per-utterance emotion classification module (utterance-to-emotion) and an emotion sequence decoder (conversation context). Given that there exists a large number of emotion classification modules available for use across diverse scenarios, this decoupling allows the decoder to focus directly on handling emotion sequence transition of interlocutors in a dialogue. This approach provides the flexibility as the decoder can wrap around any choice of emotion classification models, and the decoder itself tends to be light-weighted. For instance, Yeh et al. used IAAN as a pre-trained SER engine and designed a decoding algorithm (DED) to model the emotional dependency of intra- and inter- speakers transitions [11].

While the above works demonstrated improved recognition rates for ERC, most if not all of them, regardless of the types of approaches, have poor performances in handling cases of emotion shift. Taking DAG-ERC [10] as an example, its accuracy in the case of emotion inertia on the IEMOCAP database [12] is 74.25%, but the accuracy of the utterances with emotion shift is only 57.98%, which shows a large performance gap. Another example is the dialogical emotion decoder (DED) by Yeh et al. [11]. By re-scoring the pre-trained SER model, the prediction accuracy of the model had been improved. However, DED only improved the prediction accuracy of the utterances with emotion inertia significantly (+6.62% ACC on the IEMOCAP dataset) and had little to none improvement in the case of emotion shift (+0.23% ACC). In fact, Poria et al. pointed out even though methods for ERC has progressed much, there remains a lack of effective methods to handle the presence of emotion shift in conversation [1].

In this work, our goal is to better model the case of emotion shift by focusing on decoder design to improve ERC. In specific, we propose to use the formulation of a conditional random field (CRF) where emission score of CRF is gathered from per-utterance emotion classification module and transition score of CRF is the decoder portion. The emotion shift can then be modeled in the portion of transition score. In fact, previous work of DED [11] takes this approach, where emotion shift of DED is modeled by a simple Bernoulli distribution. Here, we propose an emotion-shift aware CRF decoder, which discriminatively learns to adjust dynamically the base transition matrix of CRF depending on whether an emotion shift is likely to occur.

We conduct experiments on English (the IEMOCAP) and Chinese (the NNIME [13]) dyadic interaction datasets. Our proposed decoder addresses the challenge of the presence of emotion shift and improves the overall speech emotion recognition performance. Specifically, it reaches 73.17% ACC on the IEMOCAP [12](2.8% better than DED), and 61.02% ACC on the NNIME [13](2.76% better than DED). In the case of emotion inertia, it improves 2.15% ACC and 2.74% ACC compared with DED on the IEMOCAP and the NNIME respectively; in the challenging case of emotion shift, it improves 4.16% ACC and 2.81% ACC compared with DED on the IEMOCAP and the NNIME respectively. Lastly, by visualize emotion-shift aware CRF's transition matrices, we observe the model's ability to automatically emphasize those frequent transitions distinctively when facing both conditions of shifting and inertia.

10.21437/Interspeech.2022-10438

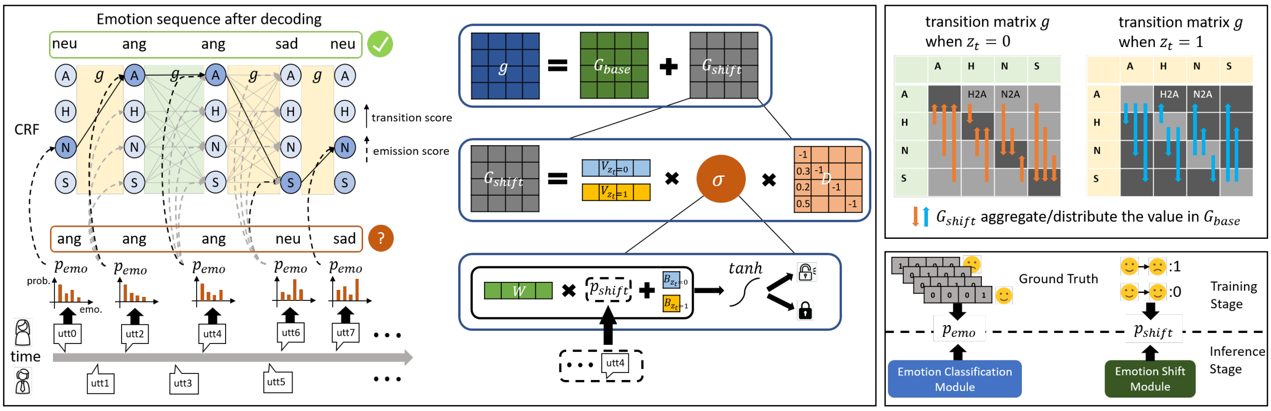

Figure 1: *An illustration of our framework of emotion sequence decoder: including an emotion-shift transition adjustment mechanism in CRF, an emotion classification module, and an emotion shift module.*

2. Methodology

2.1. Dataset Description

In this work, we evaluate our models on two dyadic interaction datasets: the IEMOCAP [12] and the NNIME [13].

IEMOCAP is one of the most widely used dataset in the field of SER. It contains 5 sessions totalling 151 dyadic conversations from 10 unique speakers. In this paper, following the conventional approached in the field (e.g., Yeh et al. [11]), we consider four categories as our emotion classification targets: anger, happiness, neutral, and sadness. Note excitement is merged into the class of happiness.

NNIME is a Chinese dyadic interaction corpus that results from the collaborative work between engineers and drama experts. It contains 22 sessions totaling 100 dyadic conversations from 43 unique speakers. Here, we consider the same four categories as the IEMOCAP: anger, happiness, neutral and sadness. Note that happiness, joy and excitement are considered together as happiness. Table 1 provides a summary of these two datasets.

2.2. Task Definition

Given a dialogue U consists of a sequence of utterances, i.e., $U = \{u_1, ..., u_T\}$ where $t \in [1, T]$, the task is to learn a model $P(y_t|x_t)$ for emotion classification, where y_t and x_t denotes the emotion and feature at utterance u_t, respectively. Conventionally, methods of ERC (e.g., [7, 8, 9, 10]) would incorporate the contextual emotion information $Y_{1:t-1} = \{y_1, ..., y_{t-1}\}$ to learn an improved model for classifying y_t. In this work, we propose to incorporate both the contextual emotion information Y and an emotion shift sequence Z as our model for decoding emotion sequence in a dialog. Our model is formulated as:

$$P(Y, Z|X) = p(y_1|x_1) \prod_{t=2}^{T} p(y_t|x_t)p(y_t, z_t|Y_{1:t-1}, Z_{1:t-1}) \tag{1}$$

where $X = \{x_1, ..., x_T\}$ is the observed features sequence. $Z = \{z_1, ..., z_T\}$ consists of z_t, where $z_t \in \{0, 1\}$ is a binary random variable. $z_t = 1$ indicates an emotion shift, i.e., the current speaker's emotion differs from his/her previous utterance, and $z_t = 0$ otherwise. In this work, we factorize $P(Y, Z|X)$ as below:

$$\begin{aligned} & p(y_t|x_t)p(y_t, z_t|Y_{1:t-1}, Z_{1:t-1}) \\ &= p(y_t|x_t)p(y_t|Z_{1:t}, Y_{1:t-1})p(z_t|x_t) \end{aligned} \tag{2}$$

where $p(y_t|x_t)$, $p(y_t|Z_{1:t}, Y_{1:t-1})$, $p(z_t|x_t)$ denote the models of emotion classification module, the emotion-shift aware contextual information, and the emotion shift module, respectively.

Table 1: *Summary of the IEMOCAP and the NNIME.*

Dataset	Ang	Hap	Neu	Sad	# Utterances	Avg. dialog length
IEMOCAP	19.9%	29.5%	30.8%	19.5%	5531	36.6
NNIME	15.9%	20.8%	54.3%	9.0%	4058	40.6

2.3. Emotion-Shift Aware CRF

In this work, we adopt the formulation of conditional random field (CRF) for implementing $P(Y, Z|X)$. CRF is a discriminatively trained sequence model that has been utilized successfully in many sequence labeling tasks, e.g., part-of-speech tagging [14, 15], named entity recognition [16, 17] and image segmentation [18]. The overall framework is shown in Fig. 1. Different from the DED decoder[11] that is based on the assumption of distance-dependent Chinese restaurant process (ddCRP) [19], we directly train a CRF model to perform sequential emotion prediction. Specifically, we propose an emotion-shift aware conditional random field (ESA CRF), i.e., a modified CRF with parameter Θ to maximize the following probability,

$$P(Y, Z|X) = \frac{\exp(\psi(X, Y, Z; \Theta))}{\sum_{\tilde{Y}} \exp(\psi(X, \tilde{Y}, Z; \Theta))}, \tag{3}$$

where X is the observed acoustic feature sequence, $\psi(X, Y, Z; \Theta)$ is a feature function that assigns a path score to the output sequence Y, and $\tilde{Y}$ denotes all possible output sequences (paths). Similar to a conventional CRF, $\psi(X, Y, Z; \Theta)$ can be defined as the sum of emission scores h and transition scores g across all time steps [20, 21]:

$$\psi(X, Y, Z; \Theta) = \sum_{t=1}^{T} h(y_t, X) + \sum_{t=2}^{T} g(y_{t-1}, y_t, z_t). \tag{4}$$

We use p_{emo} to denote $p(y_t|x_t)$ which is the emission score h. Thus, given M dialogues in our training data, the parameters Θ is updated by the loss:

$$L = \sum_{m=1}^{M} -\log P(Y^m, Z^m|X^m; \Theta). \tag{5}$$

At inference state, the optimal emotion sequence is derived by using Viterbi decoding algorithm [22] to search $Y^*, Z = \arg\max_{\tilde{Y},Z} P(\tilde{Y}, Z|X)$.

2.3.1. Emotion-Shift Adjusted Transition Matrix

The core idea of our ESA CRF decoder is learn to adjust transition scores, g, when performing sequential emotion decoding depending on whether an emotion shift has occurred. Transition score g in Eq. 4 for ESA CRF is defined as follows:

$$g(y_{t-1}, y_t, z_t) = G_{base}[y_{t-1}, y_t] + G_{shift}[y_{t-1}, y_t, z_t], \quad (6)$$

where G_{base} is a learnable base emotion transition matrix with size 4×4, e.g., the element $G_{base}[y_{t-1}, y_t]$ is the transition score from emotion y_{t-1} to y_t. Conventional CRF has a single transition matrix G_{base} that is shared across all time steps. For ESA CRF, we propose a emotion-shift gated transition adjustment mechanism. We construct another learnable transition matrix G_{shift} that would dynamically modify the value of g based on p_{shift}, i.e., $p(z_t|x_t)$. Ideally, we expect the G_{base} to enlarge the score to maintain in the same emotion state when there is no emotion shift $z_t = 0$; otherwise ($z_t = 1$) we expect the transition matrix to reduce the score of transiting to the same emotion state and distribute those scores to other potential transitions. This adjustment mechanism is implemented by the following equation:

$$G_{shift} = V_{z_t}[y_{t-1}] * \sigma(p_{shift}) * D[y_{t-1}, y_t]. \quad (7)$$

where D is an adjustment matrix of size 4×4. Following our assumption, transition matrix D is strictly defined with following rule: the diagonal values have constant coefficients of-1, and the values in every column sum up to 0, e.g., $D_{11} = -1$ and $D_{21} + D_{31} + D_{41} = +1$ for column 1. On the other hand, $V_{z_t} \in \mathbb{R}^4$ are used to re-weight the values of D for adjusting G_{base} to be emotion-shift aware. In other words, we want G_{shift} to distribute the value from diagonal to non-diagonal elements of G_{base} or aggregate the value from non-diagonal to diagonal elements of G_{base}. That is, when there is an emotion shift, we expect the transition value of non-diagonal element to increase by removing part of the original transition score from the diagonal element. Otherwise, we want the diagonal value to be large enough to keep the emotion stay at the same state by decreasing the value in non-diagonal elements. The adjustment range is controlled by a gating function σ:

$$\sigma(p_{shift}) = \tanh(W[y_{t-1}] * p_{shift} + B_{z_t}) \quad (8)$$

where $W \in \mathbb{R}^4$ and $B_{z_t} \in \mathbb{R}^1$. Based on the equation above, since the value of $\tanh$ is located in $[-1, 1]$, it acts as a gate to change the transition matrix into two modes according to z_t. V_{z_t} and B_{z_t} in Eq. 7, 8 each results in two learnable matrices depending whether $z_t = 1$ ($p_{shift} > 0.5$) or otherwise. In summary, V_{z_t}, B_{z_t}, and W in G_{shift} are all learnable weights.

2.3.2. Decoder Training and Inference

At training stage, p_{emo} is the one-hot encoded ground truth emotion label, and p_{shift} is the ground truth emotion shift binary value. At inference decoding stage, p_{emo} is the predicted probability from the pre-trained emotion classification module, and p_{shift} is the predicted probability from the pre-trained emotion shift module.

3. Experiments and Results

3.1. Experimental Setup

ESA CRF involves the use of pre-trained emotion classification and emotion shift module in the inference stage, we will first introduce the two pre-trained networks used in our experiment.

3.1.1. Emotion Classification and Emotion Shift Modules

Emotion classification module is constructed from a pre-trained SER model parameterized by Θ_{emo}, which is used to predict a probability distribution of current utterance over four emotion categories (p_{emo}). We use the exact same structure as Yeh et al. [11], i.e., an IAAN [8] model as our emotion classification module. IAAN uses attention mechanism that learns to combine the local contextual information from the current utterance (u_c), the previous utterance of the current speaker (u_p), and the previous utterance of the other speaker (u_r) as embedding used for the current utterance representation. Moreover, emotion shift estimates the speaker's emotion change between his/her current utterance and his/her previous one (p_{shift}). We use the same IAAN structure for deriving representation that learns to model the probability of emotion shift for current utterance (p_{shift}). Note that where these two modules are based on IAAN, our ESA CRF decoder is not restricted to IAAN; these two modules can be replaced with any other classifiers.

3.1.2. Comparison of ERC Decoder

We examine the performances of the following methods where p_{emo} is fixed by using the same IAAN across decoders:

* DED [11]. Baseline method: we set beam size n to 10 and 20 on the IEMOCAP and the NNIME, respectively.
* DED (bigram seq). Replace the emotion shift model and emotion assignment process in the original DED with bigram without considering speaker identity.
* CRF (seq). It is a CRF that treats a dyadic conversation as a sequence of utterances directly and does not include emotion-shift transition adjustment mechanism.
* CRF (intra). It differs from CRF (seq) by having each individual speaker in a dyadic conversation with a sequence of his/her own utterances.
* ESA CRF. Our proposed method.

3.1.3. Hyper-parameter Settings and Evaluation Metrics

In ESA CRF, one sequence of a speaker in a conversation is used for one training step (batch-size is set to be 1). The number of epoch is 70, and SGD is used as the optimizer with momentum to 0.5. The performance is evaluated using unweighted accuracy (UA) and weighted accuracy (WA). We carry out a 5-fold cross validation to evaluate the performance on both datasets where the testing conversations are strictly excluded at training. The ratio of training, validation, and testing is 3:1:1 in each fold, and the model parameters evaluated on the testing set are decided by the ones with the highest validation UA.

3.2. Results

We show our experimental results in Table 2. We find that our proposed ESA CRF achieved an overall 4 class classification performance of 73.17% WA and 74.47% UA on the IEMOCAP and 61.02% WA and 52.75% UA on the NNIME. Moreover, ESA CRF acheived a significantly higher performances compared to the DED on the IEMOCAP and the NNIME ($p <$

Table 2: *The performance of using the emotion classification module of IAAN with different decoding methods. Note that DAG [10] here uses acoustic features which are the same as IAAN [8] instead of textual features.*

Method	Overall		Emotion inertia		Emotion shift	
	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)
IEMOCAP						
DAG [10]	68.64	66.62	73.40	71.31	58.21	56.91
IAAN [8]	67.21	65.83	71.49	70.08	57.55	57.02
DED [11]	71.54	70.37	77.87	76.70	57.93	57.25
DED (bigram seq)	68.47	66.62	73.60	71.77	56.99	55.97
CRF (seq)	72.00	70.39	77.24	75.68	60.48	59.41
CRF (intra)	73.07	71.83	78.88	77.72	60.30	59.63
ESA CRF	**74.47**	**73.17**	**80.03**	**78.85**	**62.37**	**61.41**
NNIME						
IAAN [8]	49.33	52.64	52.61	56.38	40.26	41.68
DED [11]	52.56	58.26	56.68	62.93	**40.86**	44.58
DED (bigram seq)	49.97	59.54	53.85	63.86	39.18	46.91
CRF (seq)	51.14	59.31	55.01	63.82	40.42	46.13
CRF (intra)	52.62	59.93	57.22	64.52	39.67	46.52
ESA CRF	**52.75**	**61.02**	**57.34**	**65.67**	40.12	**47.39**

0.001). Compared to IAAN (without decoder module), ESA CRF has improved in both the case of emotion inertia (+8.54% UA) and emotion shift (+4.82% UA) on the IEMOCAP, which shows that having a decoder can better model both types of emotion transitions. By examining different decoding methods when given the same pre-trained emotion classification module, i.e., IAAN, we observe that ESA CRF improves both in the case of emotion inertia and shift while other decoding methods such as DED and DED (bigram seq) have more significant improvements in the case of emotion inertia only.

The use of bigram for DED leads to worse performance compared to DED on both datasets possibly due the overfitting issue. Further, we find that the variants of CRF such as CRF (seq) and CRF (intra) perform better than DED method on both of the datasets; this may attribute to the discriminative nature of CRF as opposed to generative DED. When comparing between CRFs, CRF (intra) has relative 1.07% UA and 1.48% UA improvements over CRF (seq) on the IEMOCAP and the NNIME respectively; these results show that decoding emotion sequence with long-term context for a single speaker is better since the inter-speaker influence is often short-term and is captured by the emotion classification module of IAAN already. Last but not least, we observe that in the case of emotion shift, ESA CRF achieves 2.07%, and 0.45% higher UA than CRF (intra) on the IEMOCAP and the NNIME respectively, and there is a similar improvement in the case of emotion inertia, which shows that emotion-shift transition adjustment mechanism effectively makes CRF more sensitive to transition types.

3.3. Analysis of Transitions Adjustment

In order to show the effectiveness of our transition adjusted mechanism, we compare the learned transition matrix g between the CRF (intra) and ESA CRF by visualizing the two matrices in Fig. 2(a). Based on the learned weight g, we can clearly observe the differences between CRF (intra) and ESA CRF. The matrix in CRF (intra) has larger values on the diagonal which correspond to the inertia type of emotion transition. However, since the CRF (intra) only have one transition matrix, it is less sensitive to the emotion shift case. In contrast, the g in ESA CRF evolve into two variants based on the gating function $\sigma(p_{shift})$ in Eq. 8. According to Fig. 2(a)(b), when $z_t = 0$, the matrix has large values along the diagonal; for example, the transition score of neu-to-neu increases from 1.10 to 1.28, and this leads to a relative 1.53% performance improvement in the case of neu-to-neu over CRF (intra). When $z_t = 1$, by contrast, our model automatically learns to emphasize on the frequent transition cases; for example, the transition scores of ang-to-sad and hap-to-neu increase by 0.38 and 0.27 respectively, and ESA CRF performs 5.88%, and 1.54% higher recall than CRF (intra) in the case of ang-to-sad and hap-to-neu respectively.

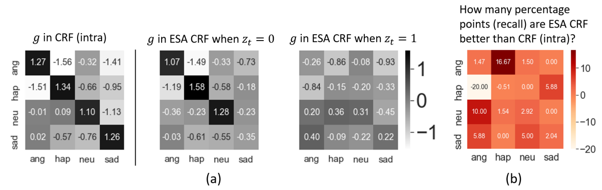

Figure 2: *(a) shows the transition matrices of CRF (intra) and ESA CRF. (b) shows the comparison of the performance of 16 emotion transitions in ESA CRF and CRF (intra). This analysis is on the IEMOCAP.*

Table 3: *ESA CRF is combined with different methods of emotion shift module (ESM). GT of ESM represents ground truth emotion shift labels (0 or 1).*

Method	ESM	Overall		Emotion inertia		Emotion shift	
	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)	UA(%)	WA(%)
IEMOCAP							
SVM	69.90	71.20	69.59	76.76	75.39	58.83	57.58
IAAN	80.18	74.47	73.17	80.03	78.85	62.37	61.41
GT	**100.00**	**78.58**	**77.62**	**82.40**	**81.34**	**70.03**	**69.91**
NNIME							
SVM	67.00	50.43	58.55	54.66	62.67	38.80	46.52
IAAN	75.75	52.75	61.02	57.34	65.67	40.12	47.39
GT	**100.00**	**56.84**	**65.55**	**60.07**	**69.61**	**47.69**	**53.68**

3.4. Comparison of Emotion Shift Modules

In order to analyze the influence of p_{shift} on the ESA CRF decoder performance, we use three methods to predict emotion shift at testing (Table 3). When we use SVM as the emotion shift module, the overall performance of ESA CRF is worse than using IAAN as the emotion shift module (-3.27% UA on the IEMOCAP and -2.32% UA on the NNIME). In view of this, the accuracy of emotion shift module plays a significant role in the decoding process. When we use ground truth emotion shift label (GT) as p_{shift}, this case represents the upper bound. We observe that all the metrics in Table 3 improve even further, especially in the case of emotion shift, which has an increase of 12.1% UA and 6.83% UA compared with DED on the IEMOCAP and the NNIME respectively. This shows that our emotion-shift aware decoder, if given powerful emotion-shift prediction module, can handle both the known emotion inertia and the challenging emotion shift when performing ERC.

4. Conclusions

In this paper, we propose a novel emotion decoding approach in ERC. We develop an emotion-shift aware CRF which includes dynamically adjustable transition matrix based on whether an emotion shift occurs. Not only does it improve the recognition accuracy of utterances in the case of emotion inertia but also the challenging case of emotion shift. Compared with other methods in handling ERC, we contribute particularly to the poor performances of emotion shift. Our method achieves the state-of-the-art performance on four emotion class UA of 74.47% on the IEMOCAP and WA of 61.02% on the NNIME. Our future research directions include investigating the robustness of ESA CRF on different pre-trained ERC models, enhancing emotion shift prediction module, and evaluating ESA CRF on multi-party dataset such as MELD [23].

5. References

[1] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, “Emotion recognition in conversation: Research challenges, datasets, and recent advances,” *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[2] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, and R. Li, “Conversational emotion recognition using self-attention mechanisms and graph neural networks.” in *INTERSPEECH*, 2020, pp. 2347–2351.

[3] M. Haugh, “Conversational interaction,” *The Cambridge handbook of pragmatics*, pp. 251–274, 2012.

[4] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic interaction patterns.* Cambridge University Press, 1995.

[5] P. Kuppens, N. B. Allen, and L. B. Sheeber, “Emotional inertia and psychological maladjustment,” *Psychological science*, vol. 21, no. 7, pp. 984–991, 2010.

[6] J. C. Veilleux, E. A. Warner, D. E. Baker, and K. D. Chamberlain, “Beliefs about emotion shift dynamically alongside momentary affect,” *Journal of Personality Disorders*, vol. 35, no. Supplement A, pp. 83–113, 2021.

[7] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, “Icon: Interactive conversational memory network for multimodal emotion detection,” in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 2594–2604.

[8] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, “An interaction-aware attention network for speech emotion recognition in spoken dialogs,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 6685–6689.

[9] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, “Dialoguernn: An attentive rnn for emotion detection in conversations,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6818–6825.

[10] W. Shen, S. Wu, Y. Yang, and X. Quan, “Directed acyclic graph network for conversational emotion recognition,” in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 1551–1560. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.123

[11] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, “A dialogical emotion decoder for speech emotion recognition in spoken dialog,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 6479–6483.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[13] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, “Nnime: The nthu-ntua chinese interactive multimodal emotion corpus,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 292–298.

[14] P. Awasthi, D. Rao, and B. Ravindran, “Part of speech tagging and chunking with hmm and crf,” *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest 2006*, 2006.

[15] Y. Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, “Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF,” in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 173–183. [Online]. Available: https://aclanthology.org/I17-1018

[16] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, “An attention-based bilstm-crf approach to document-level chemical named entity recognition,” *Bioinformatics*, vol. 34, no. 8, pp. 1381–1388, 2018.

[17] C. Dong, J. Zhang, C. Zong, M. Hattori, and H. Di, “Character-based lstm-crf with radical-level features for chinese named entity recognition,” in *NLPCC/ICCPOL*, 2016.

[18] F. Liu, G. Lin, and C. Shen, “Crf learning with cnn features for image segmentation,” *Pattern Recognition*, vol. 48, no. 10, pp. 2983–2992, 2015.

[19] D. M. Blei and P. I. Frazier, “Distance dependent chinese restaurant processes.” *Journal of Machine Learning Research*, vol. 12, no. 8, 2011.

[20] J. Morris and E. Fosler-Lussier, “Combining phonetic attributes using conditional random fields,” in *Ninth International Conference on Spoken Language Processing.* Citeseer, 2006.

[21] L. Chen and A. Moschitti, “Transfer learning for sequence labeling using source model and target data,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6260–6267.

[22] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[23] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050