Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 620–627
Marseille, 11–16 May 2020

EDA: Enriching Emotional Dialogue Acts using an Ensemble of Neural Annotators

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter
Knowledge Technology, Department of Informatics, University of Hamburg,
Vogt-Koelln-Str. 30, 22527 Hamburg, Germany
www.informatik.uni-hamburg.de/WTM/
{bothe,weber,magg,wermter}@informatik.uni-hamburg.de

Abstract

The recognition of emotion and dialogue acts enriches conversational analysis and help to build natural dialogue systems. Emotion interpretation makes us understand feelings and dialogue acts reflect the intentions and performative functions in the utterances. However, most of the textual and multi-modal conversational emotion corpora contain only emotion labels but not dialogue acts. To address this problem, we propose to use a pool of various recurrent neural models trained on a dialogue act corpus, with and without context. These neural models annotate the emotion corpora with dialogue act labels, and an ensemble annotator extracts the final dialogue act label. We annotated two accessible multi-modal emotion corpora: IEMOCAP and MELD. We analyzed the co-occurrence of emotion and dialogue act labels and discovered specific relations. For example, *Accept/Agree* dialogue acts often occur with the *Joy* emotion, *Apology* with *Sadness*, and *Thanking* with *Joy*. We make the Emotional Dialogue Acts (EDA) corpus publicly available to the research community for further study and analysis.

Keywords: Emotional Dialogue Acts Corpus, Conversational Analysis, Automated Neural Ensemble Annotation and Evaluation

1. Introduction

With the growing demand for human-computer/robot interaction systems, detecting the emotional state of the user can substantially benefit a conversational agent to respond at an appropriate emotional level. Emotion recognition in conversations has proven valuable for potential applications such as response recommendation or generation, emotion-based text-to-speech, personalization. Human emotional states can be expressed verbally and non-verbally (Ekman et al., 1987; Osgood et al., 1975). However, while building an interactive dialogue system, the interface needs dialogue acts. A typical dialogue system consists of a language understanding module which requires to determine the meaning and intention in the human input utterances (Wermter and Löchel, 1996; Berg, 2015; Ultes et al., 2017). Also, in discourse or conversational analysis, dialogue acts are the main linguistic features to consider (Bothe et al., 2018a). A dialogue act provides an intention and performative function in an utterance of the dialogue. For example, it can infer a user's intention by distinguishing *Question*, *Answer*, *Request*, *Agree/Reject*, etc. and performative functions such as *Acknowledgement*, *Conversational-opening or -closing*, *Thanking*, etc. The dialogue act information together with emotional states can be very useful for a spoken dialogue system to produce natural interaction (Ihasz and Kryssanov, 2018).

The research in emotion recognition is growing, and many datasets are available, such as text-, speech- or vision-based, and multi-modal-based emotion data. Emotion expression recognition is a challenging task, and hence multimodality is crucial (Ekman et al., 1987). However, few conversational multi-modal emotion recognition datasets are available, for example, IEMOCAP (Busso et al., 2008) or SEMAINE (McKeown et al., 2012), MELD (Poria et al., 2019). They are multi-modal dyadic conversational datasets containing audio-visual and conversational transcripts. Every utterance in these datasets is labelled with an emotion label.

In our research here, we propose an automated neural ensemble annotation process for dialogue act labelling. Several neural models are trained with the Switchboard Dialogue Act (SwDA) corpus (Godfrey et al., 1992; Jurafsky et al., 1997) and used for inferring dialogue acts on the emotion corpora. We integrate five model output labels by checking majority occurrences (most of the model labels are the same) and ranking confidence values of the models. We have annotated two potential multi-modal conversation datasets for emotion recognition: IEMOCAP (Interactive Emotional dyadic MOtion CAPture database) (Busso et al., 2008) and MELD (Multimodal EmotionLines Dataset) (Poria et al., 2019). Figure 1, shows an example of the dialogue act tags with emotion and sentiment labels from the MELD corpus and we confirmed the reliability of annotations with inter-annotator metrics. We analyzed the co-occurrences of the dialogue act and emotion labels and discovered an essential relationship between them: individual dialogue acts of the utterances show significant and useful association with corresponding emotional states. For example, the *Accept/Agree* dialogue act often occurs with the *Joy* emotion while *Reject* with *Anger*, *Acknowledgements* with *Surprise*, *Thanking* with *Joy*, and *Apology* with *Sadness*, etc. The detailed analysis of the emotional dialogue acts (EDAs) and annotated datasets are being made available at the Knowledge Technology website[1].

2. Annotation of Emotional Dialogue Acts

2.1. Data for Conversational Emotion Analysis

There are two emotion taxonomies: (1) discrete emotion categories (DEC) and (2) fined-grained dimensional ba-

[1] www.inf.uni-hamburg.de/en/inst/ab/wtm/research/corpora IEMOCAP (https://sail.usc.edu/iemocap) is available only with speaker IDs.

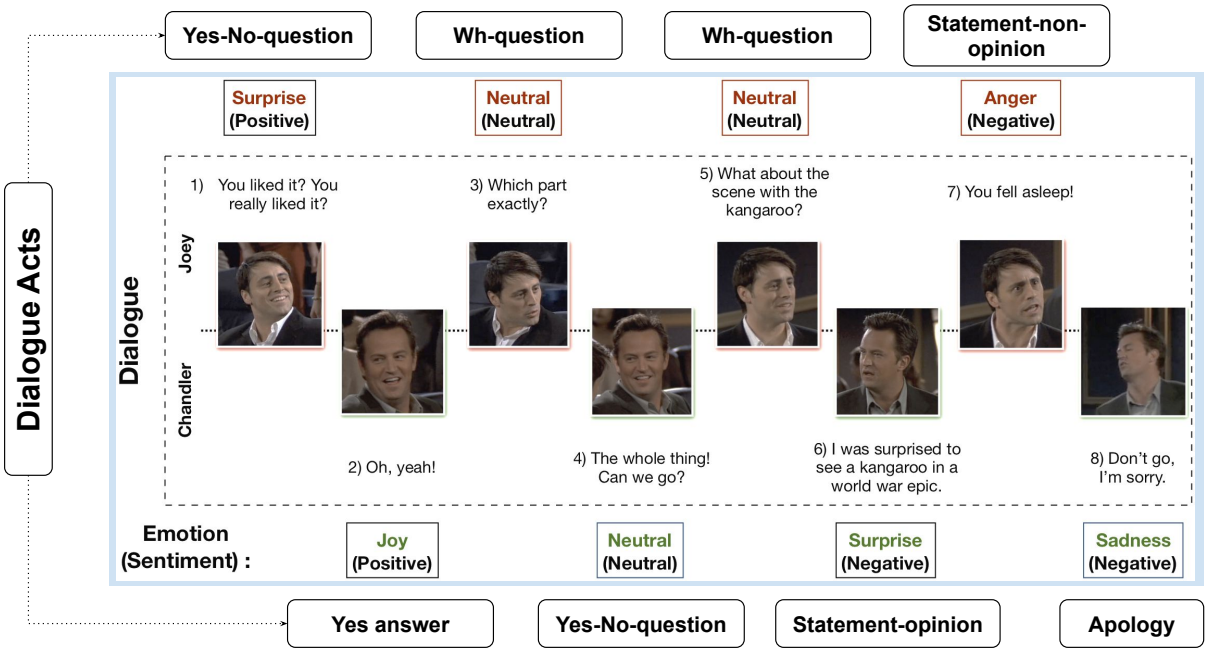


Figure 1: Emotional Dialogue Acts: Example of a dialogue from MELD representing emotions and sentiment (rectangular boxes), in our work, we add dialogue acts (rounded boxes). Image source Poria et al. (2019).

sis of emotion states (DBE). The DECs are Joy, Sadness, Fear, Surprise, Disgust, Anger and Neutral as identified by Ekman et al. (1987). The DBE of the emotion is usually elicited from two or three dimensions (Osgood et al., 1975; Russell and Mehrabian, 1977; Cowie and Cornelius, 2003). A two-dimensional model is commonly used with Valence and Arousal (also called activation), and in the three-dimensional model, the third dimension is Dominance. The IEMOCAP dataset is annotated with all DECs and two additional emotion classes, Frustration and Excited. The IEMOCAP dataset is also annotated with three DBE, that includes Valance, Arousal and Dominance (Busso et al., 2008). The MELD dataset (Poria et al., 2019), which is an evolved version of the Emotionlines dataset developed by (Chen et al., 2018), is annotated with exactly 7 DECs and sentiments (positive, negative and neutral).

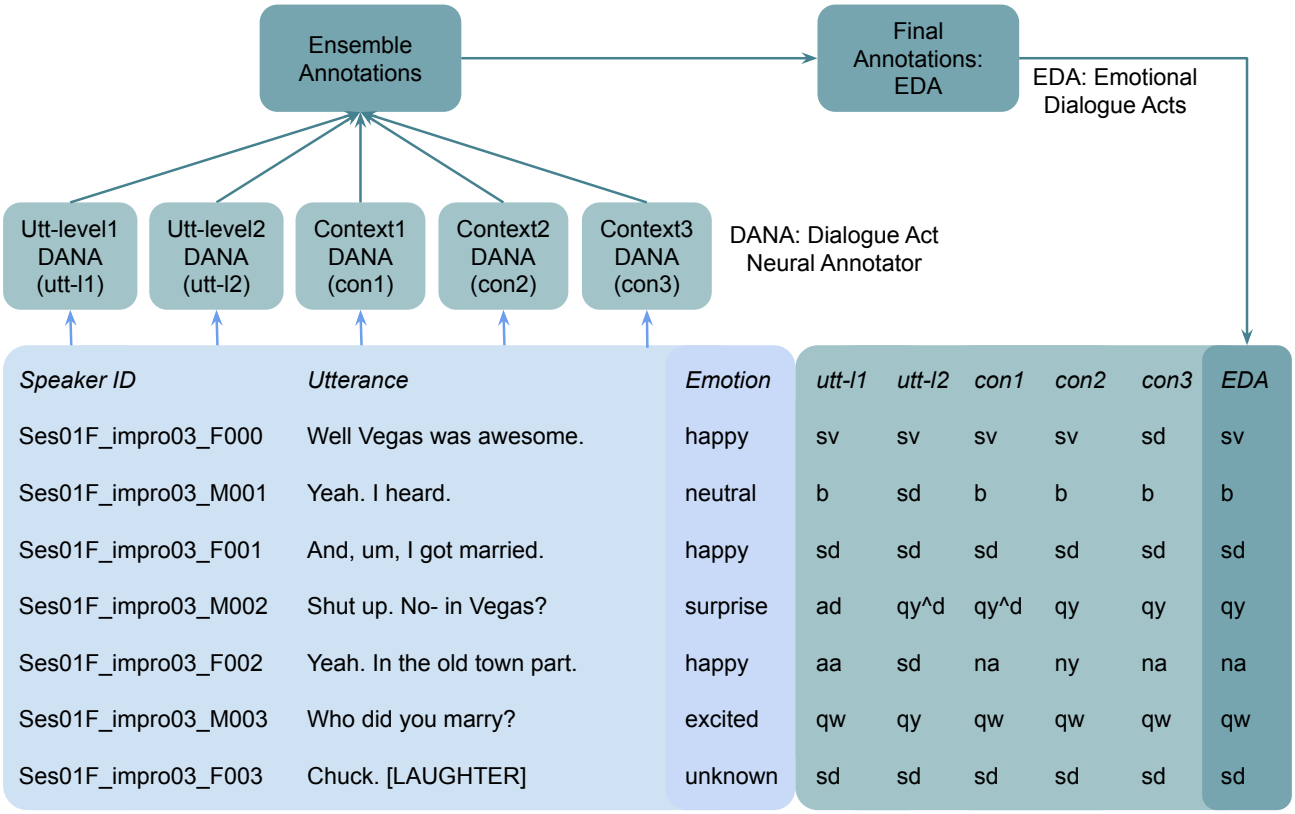
2.2. Dialogue Act Tagset and SwDA Corpus

There have been different taxonomies for dialogue acts: speech acts (Austin, 1962) refer to the utterance, not only to present information but to the action is performed. Speech acts were later modified into five classes (Assertive, Directive, Commissive, Expressive, Declarative) (Searle, 1979). There are many such standard taxonomies and schemes to annotate conversational data, and most of them follow the discourse compositionality. These schemes have proven their importance for discourse or conversational analysis (Skantze, 2007). During the increased development of dialogue systems and discourse analysis, the standard taxonomy was introduced in recent decades, called Dialogue Act Markup in Several Layers (DAMSL) tag set. According to DAMSL, each DA has a forward-looking function (such as Statement, Info-request, Question, Thanking) and a backward-looking function (such as Accept, Reject, Answer) (Allen and Core, 1997).

The DAMSL annotation includes not only the utterance-level but also segmented-utterance labelling. However, in the emotion datasets, the utterances are not segmented. As we can see in Figure 1, the first or fourth utterances are not segmented as two separate. The fourth utterance could be segmented to have two dialogue act labels, for example, a statement (*sd*) and a question (*qy*). That provides very fine-grained DA classes and follows the concept of discourse compositionality. DAMSL distinguishes wh-question *(qw)*, yes-no question *(qy)*, open-ended *(qo)*, and or-question *(qr)* classes, not just because these questions are syntactically distinct, but also because they have different forward functions (Jurafsky, 1997). For example, a *yes-no question* is more likely to get a *"yes"* answer than a wh-question *(qw)*. This gives an intuition that the context is provided by the answers (backward-looking function) with the questions (forward-looking function). For example, *qy* is used for a question that, from a discourse perspective, expects a Yes *(ny)* or No *(nn)* answer.

We have investigated the annotation method and trained our neural models with the Switchboard Dialogue Act (SwDA) Corpus (Godfrey et al., 1992; Jurafsky et al., 1997). The SwDA corpus is annotated with the DAMSL tag set, and it has been used for reporting and bench-marking state-of-the-art results in dialogue act recognition tasks (Stolcke et al., 2000; Kalchbrenner et al., 2016; Bothe et al., 2018c) which makes it ideal for our use case. The Switchboard DAMSL Coders Manual[2] has more details about the dialogue act labels (Jurafsky, 1997).

[2] https://web.stanford.edu/~jurafsky/ws97/manual.august1.html



Speaker ID	*Utterance*	*Emotion*	*utt-l1*	*utt-l2*	*con1*	*con2*	*con3*	*EDA*
Ses01F_impro03_F000	Well Vegas was awesome.	happy	sv	sv	sv	sv	sd	sv
Ses01F_impro03_M001	Yeah. I heard.	neutral	b	sd	b	b	b	b
Ses01F_impro03_F001	And, um, I got married.	happy	sd	sd	sd	sd	sd	sd
Ses01F_impro03_M002	Shut up. No- in Vegas?	surprise	ad	qy^d	qy^d	qy	qy	qy
Ses01F_impro03_F002	Yeah. In the old town part.	happy	aa	sd	na	ny	na	na
Ses01F_impro03_M003	Who did you marry?	excited	qw	qy	qw	qw	qw	qw
Ses01F_impro03_F003	Chuck. [LAUGHTER]	unknown	sd	sd	sd	sd	sd	sd

Figure 2: Setting of the annotation process of the EDAs, above example utterances (with speaker identity) and emotion labels are from IEMOCAP database.

2.3. Neural Model Annotators

We adopted the neural architectures based on Bothe et al. (2018b) where two variants are: a non-context model (classifying at utterance level) and a context model (recognizing the dialogue act of the current utterance given a few preceding utterances). From a conversational analysis using dialogue acts in Bothe et al. (2018a), we learned that the preceding two utterances contribute significantly to recognizing the dialogue act of the current utterance. Hence, we adapt this setting for the context model and create a pool of annotators using recurrent neural networks (RNNs). RNNs can model the contextual information in the sequence of words of an utterance, and the sequence of utterances of a dialogue. Each word in an utterance is represented with a word embedding vector of dimension 1024. We use the word embedding vectors from pre-trained ELMo (Embeddings from Language Models) embeddings[3] (Peters et al., 2018) as it showed promissing performance in natural language understanding tasks (Wang et al., 2018; Yang et al., 2019).

We have a pool of five neural annotators, as shown in Figure 2. Our online tool called Discourse-Wizard[4] is available to practice automated dialogue act labelling. In this tool, we use the same neural architectures but model-trained embeddings (while, in this work, we use pre-trained ELMo embeddings as they are better performant but computationally and size-wise expensive to be hosted in the online tool). The annotators are:

Utt-level-1 Dialogue Act Neural Annotator (DANA) is an utterance-level classifier that uses word embeddings (w) as an input to an RNN layer, attention mechanism (att) and computes the probability of dialogue acts (da) using the *softmax* function (see in Figure 3, dotted line utt-l1), formulated as:

$$da_t = softmax(att(RNN(w_t, w_{t-1}, ..., w_{t-m}))) \quad (1)$$

such that attention mechanism provides:

$$\sum_{n=0}^{n} a_{t-n} = 1 \quad (2)$$

This model achieved 75.13% accuracy reported in Table 1 on the SwDA corpus test set.

Context-1-DANA is a context model that uses two preceding utterances while recognizing the dialogue act of the current utterance (see context model with con1 line in Figure 3). Context-1-DANA uses a hierarchical RNN with the first RNN layer to encode the utterance from word embeddings (w) as given in equation (1) and the second RNN layer is provided with three utterances (u) (current and two preceding) composed from the first layer followed by the attention mechanism (a). Finally, the *softmax* function is used to compute the probability distribution, which is formulated as:

$$da_t = softmax(att(RNN(u_t, u_{t-1}, u_{t-2}))) \quad (3)$$

[3] https://allennlp.org/elmo

[4] https://secure-robots.eu/fellows/bothe/discourse-wizard-demo/

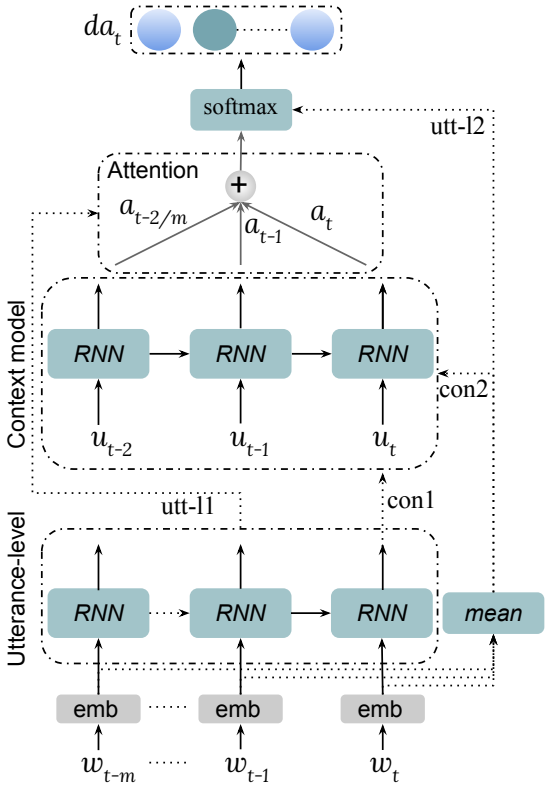


Figure 3: Recurrent neural attention architecture with the utterance-level and context-based models.

where u_t is derived from $RNN(w_t, w_{t-1}, ..., w_{t-m})$. This model achieved 77.55% accuracy on the SwDA corpus test set see Table 1, it is highest performant model among the five annotators.

Utt-level-2-DANA is another utterance-level classifier which takes an average of the word embeddings in the input utterance and uses a feedforward neural network hidden layer (see the utt-l2 line in Figure 3, where $mean$ passed to $softmax$ directly). Similar to the previous model, it computes the probability of dialogue acts using the *softmax* function. This model achieved 72.59% accuracy on the test set of the SwDA corpus (see Table 1).

Context-2-DANA is another context model that uses three utterances similar to the Context-1-DANA model. However, the utterances are composed of the mean of the word embeddings over each utterance, similar to the Utt-level-2-DANA model ($mean$ passed to context model in Figure 3 with con2 line). Hence, the Context-2-DANA model is composed of one RNN layer with three input vectors, finally topped with the *softmax* function for computing the probability distribution of the dialogue acts. This model achieved 75.97% accuracy on the test set of the SwDA corpus (see Table 1).

Context-3-DANA is a context model that uses three utterances similar to the previous context models. However, the utterance representations combine both features from the Context-1 and Context-2 models (con1 and con2 together in Figure 3). Hence, the Context-3-DANA model combines features of almost all the previous four models to provide the recognition of the dialogue acts. This model achieves 74.91% accuracy on the SwDA corpus test set (in Table 1).

Models	Accuracy	SC
Utt-level-1 mdoel	0.751	0.815
Context-1 mdoel	0.775	0.829
Utt-level-2 mdoel	0.726	0.806
Context-2 mdoel	0.759	0.823
Context-3 mdoel	0.749	0.820
Ensemble mdoel	0.778	0.822

Table 1: Baseline validation with the SwDA test dataset. SC: Spearman Correlation between prediction of model and ground truth.

2.4. Ensemble of Neural Annotators

As a baseline to verify the ensemble logic, we use the SwDA test dataset where we know the ground truth labels. Table 1 shows the accuracy and Spearman correlation between the prediction of the model and the ground truth. The ensemble model logic is configured in a way that it achieves an accuracy similar to or better than one of the neural annotators. As can be seen in Table 1, the ensemble model achieves equivalent or a little bit better accuracy to the Context-1 model. It is shown that the ensemble annotator performs well on the state of the art test data. These results are also supported by the correlation scores of the respective models. Hence, the configuration for the ensemble model that achieved the accuracy for the SwDA test dataset is explained in the following paragraph.

First preference is given to the labels that are perfectly matching in all the neural annotators. In Table 2, we can see that both datasets have about 40% of exactly matching labels over all the models (AM). Then priority is given to the context-based models to check if the label in all context models is matching perfectly. In case two out of three context models are correct, then it is being checked if that label is also produced by at least one of the non-context models. Then, we allow labels to rely on these at least two context models. As a result, about 50% of the labels are taken based on the context models (CM). When none of the context models is producing the same results, then we rank the labels with their respective confidence values produced as a probability distribution using the $softmax$ function. The labels are sorted in descending order according to confidence values. Then we check if the first three (case when one context model and both non-context models produce

Stats	AM	CM	BM	NM
IEMOCAP	43.73	50.21	1.18	4.88
MELD	37.07	51.56	2.20	9.17

Table 2: Annotations Statistics of EDAs - AM: All Absolute Match (in %), CM: Context-based Models Absolute Match (in %, matched all context models or at least two context models matched with one non-context model), BM: Based-on Confidence Ranking, and NM: No Match (in %) (these labeled as 'xx': determined in EDAs).

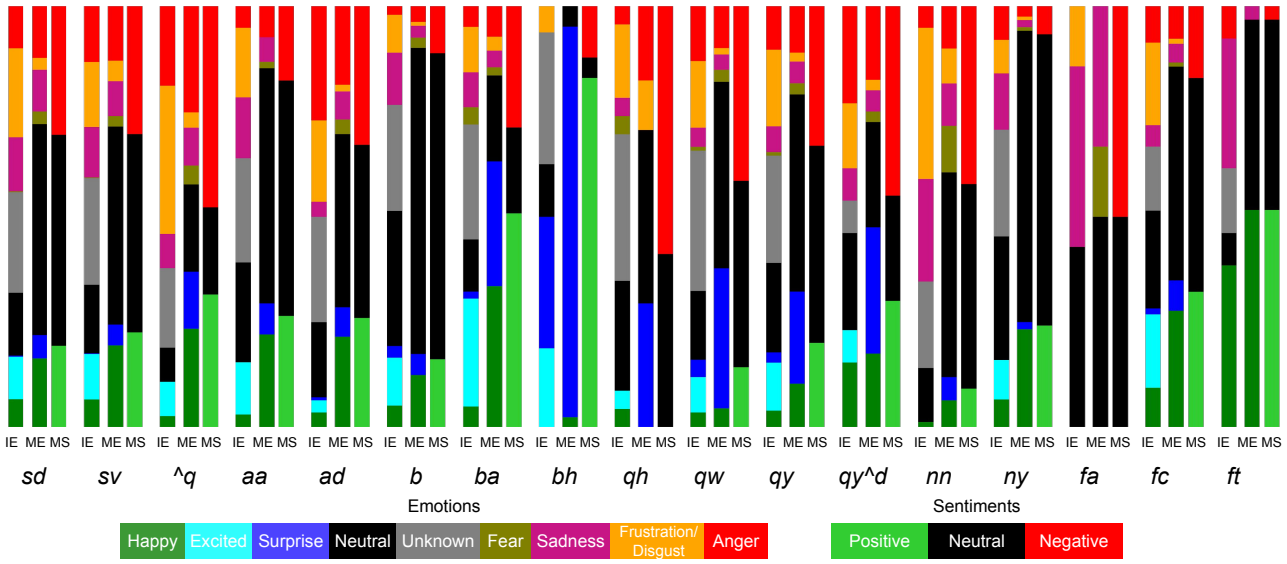


Figure 4: EDAs: Visualizing co-occurrence of utterances with respect to emotion states in the particular dialogue acts (only major and significant are shown here). IE: IEMOCAP, ME: MELD Emotion and MS: MELD Sentiment.

DA	Dialogue Act names	IEMO	MELD
sd	Statement-non-opinion	43.97	41.63
sv	Statement-opinion	19.93	09.34
qy	Yes-No-Question	10.3	12.39
qw	Wh-Question	7.26	6.08
b	Acknowledge (Backchannel)	2.89	2.35
ad	Action-directive	1.39	2.31
fc	Conventional-closing	1.37	3.76
ba	Appreciation or Assessment	1.21	3.72
aa	Agree or Accept	0.97	0.50
nn	No-Answer	0.78	0.80
ny	Yes-Answer	0.75	0.88
br	Signal-non-understanding	0.47	1.13
^q	Quotation	0.37	0.81
na	Affirmative non-yes answers	0.25	0.34
qh	Rhetorical-Question	0.23	0.12
bh	Rhetorical Backchannel	0.16	0.30
h	Hedge	0.15	0.02
qo	Open-question	0.14	0.10
ft	Thanking	0.13	0.23
qy^d	Declarative Yes-No-Question	0.13	0.29
bf	Reformulate	0.12	0.19
fp	Conventional-opening	0.12	1.19
fa	Apology	0.07	0.04
fo	Other Forward Function	0.02	0.05
Total	number of utterances	10039	13708

Table 3: Number of utterances per DA in the respective datasets. All values are in percentages (%) of the total number of utterances. IEMO is for IEMOCAP.

the same label) or at least two labels are matching, then we allow to pick that one. There are about 1% in IEMOCAP and 2% in MELD (BM).

Finally, when none the above conditions are fulfilled, we leave out the label with an unknown category. This unknown category of the determined dialogue act is labelled with 'xx' in the final annotations, and they are about 5% in IEMOCAP and 9% in MELD (NM). The statistics[5] of the EDAs is reported in Table 3 for both corpora. Total utterances in MELD includes training, validation and test datasets[6].

Metrics	α	k	SCC
IEMOCAP	0.553	0.556	0.636
MELD	0.494	0.502	0.585

Table 4: Annotations Metrics of EDAs - α: Krippendorff's Alpha coefficient, k: Fleiss' Kappa score, and SCC: Spearman Correlation between Context-based Models.

2.5. Reliability of Neural Annotators

The pool of neural annotators provides a fair range of annotations, and we checked the reliability with the following metrics (McHugh, 2012). Krippendorff's Alpha (α) is a reliability coefficient developed to measure the agreement among observers, annotators, and raters, and is often used in emotion annotation (Krippendorff, 1970; Wood et al., 2018). We apply it on the five neural annotators at the nominal level of measurement of dialogue act categories. α is computed as follows:

$$\alpha = 1 - \frac{D_o}{D_e} \quad (4)$$

where D_o is the observed disagreement and D_e is the disagreement that is expected by chance. $\alpha = 1$ means all annotators produce the same label, while $\alpha = 0$ would mean none agreed on any label. As we can see in Table 4, both

[5]We are working on improving the ensemble annotation logic; hence the updated statistics will be available at the link given on the first page where corpora are available.

[6]https://affective-meld.github.io/

EDAs	Utterances	Emotion	Sentiment
Quotation (^q)	Not after this!	anger	negative
	Ross, I am a human doodle!!	anger	negative
	No, you can't let this stop you from getting massages!	sadness	negative
	Oh hey! You got my parent's gift!	joy	positive
Action-Directive (ad)	And stop using my name!	anger	negative
	Oh, let's not tell this story.	sadness	negative
	Check it out, he's winning!	surprise	positive
	Yep! Grab a plate.	joy	positive
Acknowledgement/Backchannel (b)	Oh yeah, sure.	neutral	neutral
Appreciation Backchannel (ba)	Great.	joy	positive
Rhetorical Backchannel (bh)	Oh really?!	surprise	positive
Rhetorical Question (qh)	Oh, why is it unfair?	surprise	negative
Wh-Question (qw)	What are you doing?	surprise	negative
	How are you?	neutral	neutral
Yes-No Question (qy)	Did you just make that up?	surprise	positive
Declarative Yes-No Question (qy^d)	Can't you figure that out based on my date of birth?	anger	negative
No-Answer (nn)	No!	disgust	negative
Yes-Answer (ny)	Yeah!	joy	positive
Determined EDAs (xx)			
1. (P-DA b) b, b, ba, fc, b	Yeah, sure!	neutral	neutral
2. (P-DA sd) sv, aa, bf, sv, nn	No way!	surprise	negative
3. (P-DA qy) aa, aa, ng, ny, nn	Um-mm, yeah right!	surprise	negative
4. (P-DA qy) aa, ar, ^q, ^h, nn	Oh no-no-no, give me some specifics.	anger	negative
5. (P-DA fc) fc, sd, fc, sd, fp	I'm so sorry!	sadness	negative

Table 5: Examples of EDAs with annotation from the MELD dataset. Emotion and sentiment labels are given in the dataset, while our ensemble of models determines EDAs. P-DA: previous utterance dialogue act.

datasets IEMOCAP and MELD produce significant inter-neural annotator agreement, 0.553 and 0.494, respectively. However, it is a well-known problem with Kappa (Powers, 2012), that dialogue acts are highly subjective and contain the unbalanced number of samples per category; still, we reach these average scores. Hence, we decided to add one more inter-annotator metric below.

A very popular inter-annotator metric is Fleiss' Kappa score (Fleiss, 1971), also reported in Table 4, which determines consistency in the ratings. The kappa k can be defined as,

$$k = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{5}$$

where the denominator $1 - \bar{P}_e$ elicits the degree of agreement that is attainable above chance, and the numerator $\bar{P} - \bar{P}_e$ provides the degree of the agreement actually achieved above chance. Hence, $k = 1$ if the raters agree completely, and $k = 0$ when they do not reach any agreement. We got 0.556 and 0.502 for IEOMOCAP and MELD, respectively, with our five neural annotators. This indicates that the annotators are labelling the dialogue acts reliably and consistently. We also report the Spearman's correlation between context-based models (Context-1 and Context-2), and we find a strong correlation between them (Table 4). While using the labels, we checked the absolute match between all context-based models and found that their strong correlation indicates their robustness.

3. EDAs Analysis

We can see emotional dialogue act co-occurrences with respect to emotion labels in Figure 4 for both datasets. There are sets of three bars per dialogue act in the figure, the first and second bar represents emotion labels of IEMOCAP (IE) and MELD (ME), and the third bar is for MELD sentiment (MS) labels. MELD emotion and sentiment statistics are compelling as they are strongly correlated to each other. The bars contain the normalized number of utterances for emotion labels concerning the total number of utterances for that particular dialogue act category. The statements without-opinion (*sd*) and with-opinion (*sv*) contain utterances with almost all emotions. Many neutral utterances are spanning over all the dialogue acts.

Quotation (*^q*) dialogue acts, on the other hand, are mostly used with 'Anger' and 'Frustration' (in case of IEMOCAP), but some utterances with 'Joy' or 'Sadness' as well (see examples in Table 5). Action Directive (*ad*) dialogue act utterances, which are usually orders, frequently occur with 'Anger' or 'Frustration' although many also with the 'Happy' emotion in case of the MELD dataset. Acknowledgements (*b*) are mostly used with positive or neutral sentiment, however, Appreciation (*ba*) and Rhetorical (*bh*) backchannels often occur with a greater number in 'Surprise', 'Joy' and/or with 'Excited' (in case of IEMOCAP). Questions (*qh, qw, qy and qy^d*) are mostly asked with emotions 'Surprise', 'Excited', 'Frustration' or 'Disgust' (in case of MELD), and many are neutral. No-answers (*nn*) are mostly 'Sad' or 'Frustrated' as compared to yes-

Utterances	Emotion	Annotators	EDA
I'm sorry.	sadness	ba,sd,fc,fc,fc	fc
Dude, I am sorry about what I said!	sadness	sd,fa,^q,sd,sd	sd
Sorry, Pheebs.	sadness	fc,fa,ad,fa,ad	ad
I am so sorry...	sadness	sd,sd,^q,sd,sd	sd
Thank you.	neutral	fc,fc,ba,ft,ba	ba
Thank you we're so excited.	joy	fc,sd,fc,fc,fc	fc
Nice, thank you.	joy	fc,fc,ba,ft,ba	ba

Table 6: Examples of wrongly determined (or confused) EDAs with annotation from the MELD dataset.

answers (*ny*). Forward-functions such as Apology (*fa*) are mostly used with 'Sadness' whereas Thanking (*ft*) and Conventional-closing or -opening (*fc* or *fp*) are usually with 'Joy' or 'Excited'.

We also noticed that both datasets exhibit a similar relation between dialogue act and emotion. The dialogue act annotation is based on the given transcripts; however, the emotional expressions are better perceived with audio or video (Busso et al., 2008; Lakomkin et al., 2019). We report some examples where we mark the utterances with a determined label ('xx') in the last row of Table 5. They are left out from the final annotation (labeled as determined EDA 'xx') because of not fulfilling the conditions explained in Section 2.4. It is also interesting to see the previous utterance dialogue acts (P-DA) of those skipped utterances, and the sequence of the labels can be followed from Figure 2 (utt-l1, utt-l2, con1, con2, con3).

In the first example, the previous utterance was *b*, and three DANA models produced labels of the current utterance as *b*, but it is skipped because the confidence values were not sufficient to bring it as a final label. The second utterance can be challenging even for humans to decide with any of the dialogue acts. However, the third and fourth utterances are followed by a yes-no question (*qy*), and hence, we can see in the third example, that context models tried their best to at least perceive it as an answer (*ng, ny, nn*).

The last utterance, "I'm so sorry!", has different results by all the five annotators. Similar apology phrases are mostly found with 'Sadness' emotion label, and the correct dialogue act is Apology (*fa*). However, they are placed either in the *sd* or in *ba* dialogue act category. This mostly occurs due to less number of examples in the dialogue act categories like *fa* or *ft*. See Table 6, where the EDAs are either wrongly determined or confused by all the annotators. It is essential that the context-based models are looking into the previous utterances; hence, the utterance "Thank you." can be treated as backchannel acknowledgement (*ba*). Hence, we believe that with human annotator's help, those labels of the utterances can be corrected with minimal efforts.

4. Conclusion and Future Work

In this work, we presented a method to extend conversational multi-modal emotion datasets with dialogue act labels. The ensemble model of the neural annotators was tested on the Switchboard Dialogue Acts corpus test set to validate its performance. We successfully annotated two well-established emotion datasets: IEMOCAP and MELD, which we labelled with dialogue acts and made them publicly available for further study and research. As a first insight, we found that many of the dialogue acts and emotion labels follow certain relations. These relations can be useful to learn about the emotional behaviours with dialogue acts to build a natural dialogue system and for more in-depth conversational analysis. The association between dialogue act and emotion labels is highly subjective. However, the conversational agent might benefit in generating an appropriate response when considering both emotional states and dialogue acts in the utterances.

In future work, we foresee the human in the loop for the annotation process along with a pool of automated neural annotators. Robust annotations can be achieved with minimal human effort and supervision, for example, observing and correcting the final labels produced by ensemble output labels from the neural annotators. The human-annotator might also help to achieve segmented-utterance labelling of the dialogue acts. We also plan to use these corpora for conversational analysis to infer interactive behaviours of the emotional states with respect to dialogue acts. In our recent work, where we used dialogue acts to build a dialogue system for a social robot, we find this study and datasets very helpful. For example, we can extend our robotic conversational system to consider emotion as an added linguistic feature to produce a more natural interaction.

5. Acknowledgements

We would like to acknowledge funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant agreement No 642667 (SECURE).

6. Bibliographical References

Allen, J. and Core, M. (1997). Draft of DAMSL: Dialogue Act Markup in Several Layers. *Carnegie Mellon University*.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.

Berg, M. M. (2015). NADIA: A Simplified Approach Towards the Development of Natural Dialogue Systems. In *International Conference on Applications of Natural Language to Information Systems*, pages 144–150. Springer.

Bothe, C., Magg, S., Weber, C., and Wermter, S. (2018a). Conversational Analysis using Utterance-level Attention-based Bidirectional Recurrent Neural Networks. In *Proceedings of the International Conference INTERSPEECH 2018*, pages 996–1000. International Speech Communication Association (ISCA).

Bothe, C., Magg, S., Weber, C., and Wermter, S. (2018b). Discourse-Wizard: Discovering Deep Discourse Structure in your Conversation with RNNs. *preprint arXiv:1806.11420*.

Bothe, C., Weber, C., Magg, S., and Wermter, S. (2018c). A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks. In *Proceedings of the Eleventh International Conference on*

Language Resources and Evaluation, LREC 2018, pages 1952–1957. European Language Resources Association (ERLA).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Huang, T.-H., and Ku, L.-W. (2018). EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, pages 1597–1601.

Cowie, R. and Cornelius, R. R. (2003). Describing the Emotional States That Are Expressed in Speech. *Speech Communication*, 40(1-2):5–32.

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., et al. (1987). Universals and Cultural Differences in the Judgments of Facial Expressions of Emotion. *Journal of Personality and Social Psychology*, 53(4):712–717.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 517–520.

Ihasz, P. L. and Kryssanov, V. (2018). Emotions and Intentions Mediated with Dialogue Acts. In *Proceedings of the 5th International Conference on Business and Industrial Research (ICBIR)*, pages 125–130. IEEE.

Jurafsky, D., Shriberg, E., and Biasca, D. (1997). Switchboard Dialog Act Corpus. Technical report, International Computer Science Inst. Berkeley CA.

Jurafsky, D. (1997). Switchboard SWBD-DAMSL Shallow-Discourse-Function Annotation Coders Manual, draft 13. *Technical Report 97-01, University of Colorado Institute of Cognitive Science*, pages 225–233.

Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A. v. d., Graves, A., and Kavukcuoglu, K. (2016). Neural Machine Translation in Linear Time. *arXiv:1610.10099*.

Krippendorff, K. (1970). Estimating the Reliability, Systematic Error and Random Error of Interval Data. *Educational and Psychological Measurement*, 30(1):61–70.

Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., and Wermter, S. (2019). Incorporating End-to-End Speech Recognition Models for Sentiment Analysis. In *International Conference on Robotics and Automation (ICRA)*, pages 7976–7982. IEEE.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2012). The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Osgood, C. E., May, W. H., Miron, M. S., and Miron, M. S. (1975). *Cross-cultural Universals of Affective Meaning*, volume 1. University of Illinois Press.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. *North American Chapter of the Association for Computational Linguistics (NAACL 2018)*, pages 2227–2237.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.

Powers, D. M. (2012). The Problem with Kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 345–355. Association for Computational Linguistics.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality*, 11(3):273–294.

Searle, J. R. (1979). *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press.

Skantze, G. (2007). Error Handling in Spoken Dialogue Systems-Managing Uncertainty, Grounding and Miscommunication: Chapter 2, Spoken Dialogue Systems. *KTH Computer Science and Communication*.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., and Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Ultes, S., Rojas Barahona, L. M., Su, P.-H., Vandyke, D., Kim, D., Casanueva, I., Budzianowski, P., Mrkšić, N., Wen, T.-H., Gasic, M., and Young, S. (2017). PyDial: A Multi-domain Statistical Dialogue System Toolkit. In *Proceedings of the ACL 2017, System Demonstrations*, pages 73–78. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

Wermter, S. and Löchel, M. (1996). Learning dialog act processing. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2, pages 740–745. Association for Computational Linguistics.

Wood, I., McCrae, J. P., Andryushechkin, V., and Buitelaar, P. (2018). A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, pages 1197–1202.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in neural information processing systems*, pages 5754–5764.