INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Detecting Topic-Oriented Speaker Stance in Conversational Speech

Catherine Lai,[1] *Beatrice Alex,*[1,2] *Johanna D. Moore,*[1] *Leimin Tian,*[3] *Tatsuro Hori,*[4] *Gianpiero Francesca*[5]

[1]School of Informatics, University of Edinburgh, Edinburgh, UK
[2]Edinburgh Futures Institute, School of Literatures, Languages and Cultures, University of Edinburgh, Edinburgh, UK
[3]Computer Human Interaction & Creativity, Monash University, Melbourne, Australia
[4]Toyota Motor Corporation, Tokyo, Japan
[5]Toyota Motor Europe, Brussels, Belgium

{c.lai,jmoore,balex}@ed.ac.uk, leimin.tian@monash.edu,
tatsuro_hori@mail.toyota.co.jp, gianpiero.francesca@toyota-europe.com

Abstract

Being able to detect topics and speaker stances in conversations is a key requirement for developing spoken language understanding systems that are personalized and adaptive. In this work, we explore how topic-oriented speaker stance is expressed in conversational speech. To do this, we present a new set of topic and stance annotations of the CallHome corpus of spontaneous dialogues. Specifically, we focus on six stances—positivity, certainty, surprise, amusement, interest, and comfort—which are useful for characterizing important aspects of a conversation, such as whether a conversation is going well or not. Based on this, we investigate the use of neural network models for automatically detecting speaker stance from speech in multi-turn, multi-speaker contexts. In particular, we examine how performance changes depending on how input feature representations are constructed and how this is related to dialogue structure. Our experiments show that incorporating both lexical and acoustic features is beneficial for stance detection. However, we observe variation in whether using hierarchical models for encoding lexical and acoustic information improves performance, suggesting that some aspects of speaker stance are expressed more locally than others. Overall, our findings highlight the importance of modelling interaction dynamics and non-lexical content for stance detection.

Index Terms: spoken language understanding, affective computing, stance, computational paralinguistics, spoken dialogue.

1. Introduction

Access to speech-based interfaces is becoming increasingly important for navigating the modern world, especially for people with reduced physical mobility. Such interfaces could be particularly useful for assistive robots in building long-term engagement strategies with humans, making their support services more personalized and satisfying. In order to achieve this, speech technologies with robust language understanding capabilities are needed. At a minimum, speech technologies should be able to detect what people are talking about, i.e. the topic of conversation. However, to enable personalized and engaging interactions, we also need to detect how speakers feel about specific parts of a conversation, i.e. speaker stance. For instance, a speaker's stance can be used to understand whether a conversation is going well (or not) from each speaker's perspective.

In this paper, we investigate automatic detection of speaker stance in conversational speech scenarios. Following [1], we view a speaker stance as an evaluation of an object through which a speaker positions themselves with respect to other speakers in the dialogue. In our case, the object of evaluation is the topic of conversation itself. Understanding a speaker's position is a powerful cue for guiding interaction strategies. For example, if a speaker enjoys talking about politics, we may want to extend the discussion, while if it makes them feel uncomfortable, we may want to avoid the topic. Similarly, it would be useful be able to detect if speakers are uncertain or surprised when performing new tasks or learning new information [2].

An ability to automatically detect speaker stance would be useful for many language understanding and affective computing applications. In fact, a considerable amount of work has been done in detecting political stances from text, as a distinct task from sentiment analysis [3, 4, 5]. However, relatively little work has been done on detecting stance in speech compared to automatic emotion recognition [6, 7, 8] or sentiment analysis [9, 10, 11]. As with emotion recognition, in order to make this problem tractable, we need to constrain the types of stances we detect. This approach is the basis of previous studies which focus on quite disparate types of stance. For example, [12] identify 14 non-topic stances in news content that may aid in language technologies for disaster relief (e.g. whether a news item has bad implications, is controversial, is locally oriented). In contrast, [13, 14, 15] focus on detecting stance polarity and strength related to stance acts in task-oriented speech (e.g. option-offering vs opinion-soliciting), rather than capturing specific types of evaluations. Despite their different goals, these studies have identified consistent phonetic/prosodic correlates of stance related categories in speech, supporting the idea that non-lexical features are a rich indicators of speaker stance.

To investigate how stance is expressed in conversational dialogues, we present a new set of topic-oriented stance annotations of the CallHome English corpus. We focus on six different stance dimensions: positivity, certainty, surprise, amusement, interest, and comfort. These dimensions are more specific than what is usually used in sentiment analysis or emotion recognition, giving us a more nuanced picture of how positive affect is expressed in conversation. For example, a speaker may find it very amusing/enjoyable to talk about how much they hate their work. We expect this will, in turn, give us a clearer, more targeted view of how lexical and acoustic aspects of speech convey a speaker's affective state.

In this paper, we present experiments on automatically predicting speaker stances over multi-turn topic segments. In Sec-

http://dx.doi.org/10.21437/Interspeech.2019-2632

tion 2, we describe our annotation procedure. In Section 3, we describe our experimental setup for exploring different automatic stance detection models. We would expect our stances to be expressed differently across topic segments. For example, we would expect expressions of surprise to occur on specific turns, while comfort or interest may only be detectable over multiple turns. To investigate this, we evaluate neural network architectures with varying levels of hierarchical structure (Section 4). We compare our proposed hierarchical approaches with DialogueRNN [16], a state-of-the-art neural architecture for emotion recognition. Finally, we conclude with a discussion of our results and ideas for future work (Section 5).

2. Topic and Stance Annotations

2.1. The CallHome English Corpus

Our long term goal is to enable robot companions to interact with their human counterparts in a personal, conversational manner that maintains good inter-personal rapport. With this in mind, we chose to analyze the the CallHome English corpus (LDC97S42, LDC97T14). This includes 120 unscripted dialogues between native speakers of English. Participants spoke over the telephone to a person of their choice (generally family members or close friends) for around 30 minutes. Around 10 minutes of each conversation was manually transcribed and segmented into time stamped speaker turns. The transcripts were additionally marked up for named entities (i.e., names of people/places). Since these dialogues are between people who know each other well, the type of language used is personal and expressive, and the discussions covers a wide range of topics related to daily life. Thus, it provides a good fit for our goals.

2.2. Identification of Potential Topics

To detect topic segments, we first identify potential topics for each conversation, relating to named entities, frequent nouns, and life events, as well as broad topics induced using probabilistic topic modelling techniques. We extracted named entities from the transcript text markup. We identify nouns that occur 5 or more times (e.g. ‘baby’, ‘house’, ‘university’), and mentions of words relating to major life events if they occur at all in a conversation (e.g., ‘wedding’, ‘birth’, ‘death’, etc).

To extract broad topic categories, we use Latent Dirichlet Allocation (LDA) [17] to obtain a topic model using the Mallet toolkit [18]. To improve the coherence and robustness of the learned topics, we augment the CallHome transcripts with the much larger Fisher Corpus (LDC2004T19) which includes 11699 telephone conversations between strangers. While there is a lot of topical overlap between the corpora, the nature of the conversations is quite different (Fisher speakers displaying much less rapport). We fit a 100 topic LDA model on the combined data set, removing stop words, discourse markers and other high document frequency words. We also remove named entities in the CallHome data to avoid associating common names with specific topics. We inspected the high probability words associated with learned topics to manually identify 35 broad topics. The first 16 topics broadly cover the basic aspects of day-to-day life (e.g. scheduling, work, family), while the rest address more specific speaker interests (e.g. sport, music, books). We set a threshold for associating a broad topic with a conversation (topic probability > 0.01). In the end, the CallHome conversations were associated with 32 topic keywords on average, with a high variance between conversations (standard deviation of 11)

Positive/Negative: Does the speaker appear to like or dislike the person/place/event being discussed?

Surprised/Unsurprised: Does the information being discussed seem new to speaker or is it something obvious and to be expected?

Certain/Uncertain: Does the speaker sound sure of themselves or like they are lacking in knowledge?

Interested/Uninterested: Is this something the speaker cares about or are they bored?

Comfortable/Uncomfortable: Is this something the speaker likes talking about or would they rather not speak about it?

Amused/Annoyed: Does the speaker seem to be finding the discussion funny/enjoyable or are they annoyed/irritated by it?

Figure 1: *Speaker stance annotation dimensions.*

2.3. Topic Segment Annotations

Topic Segment annotations were carried out via Amazon Mechanical Turk (AMT) using a modified version of the Brat annotation tool [19]. For each Human Intelligence Task (HIT), annotators were given four potential topic keywords for a specific CallHome conversation. They then listened to and read the transcribed portion of the conversation and marked up all (multi-turn) segments of the transcript where the speakers talked about the given keyword in that conversation (if any). This approach allowed us to induce a non-linear topic structure on the conversations, while reducing the cognitive load on annotators and parallellizing the annotation process. Topic segment identification of the complete CallHome corpus was completed by 12 AMT workers. The standard CallHome development and test sets (40 conversations) were annotated two additional times to get an idea of inter-annotator agreement. We observe Krippendorf’s α [20] of 0.61 over all topic keywords, which indicates that the annotators could do the task reasonably reliably.

2.4. Stance Annotations

We obtained speaker stance annotations for 4290 topic segments via AMT. This included all manually identified topic segments which had a duration between 5 and 60 seconds, contained at least 2 speaker turns, and involved both speakers. These constraints were put in place to make sure annotators had enough context to gauge the stances of both conversational participants, but weren’t so long that speaker stances were likely to change significantly over the segment. We use the union of overlapping topic segments from different annotators for the multiply annotated dev and test set conversations.

In each HIT, annotators listened to the audio and read the transcript for a given topic segment. They then rated each speaker’s stance in the clip along the dimensions shown in Figure 1. These stances were chosen as important dimensions for understanding speaker rapport after preliminary analysis of the data. Ratings were selected from a 5 point Likert scale, e.g. very positive (2), positive (1), neutral (0), negative (-1), very negative (-2). Each topic segment was annotated three times, resulting in 12870 HITs completed by 61 annotators.

In terms of inter-annotator agreement, Krippendorff’s α over the 5-point stance ratings was relatively low: $\alpha = 0.18$. 36% of the segments obtained the same rating from all annotators when we collapse ratings to three ‘sign’ classes (i.e. positive: $r > 0$, neutral: $r = 0$, or negative: $r < 0$). However, we found that most disagreements were between either neutral and positive ratings or neutral and negative ratings. In fact, only 4% of segments received clashing positive and negative ratings. This was deemed to be sufficient agreement to use this data to investigate how our stances are expressed in everyday speech. We use majority vote to determine gold standard stance ratings in the experiments that follow.

3. Experimental Setup

3.1. Task Definition and Evaluation

Our immediate goal is to see whether automatic stance detection on our data set is feasible. So, we cast our task as a binary classification task (class 1: segments with ratings > 0, class 0: ratings ≤ 0) and leave experiments using the fuller rating scales for future work. We build separate models for each of our six stance types, thus each classification can be interpreted as querying whether a specific speaker is *amused, certain, comfortable, interested, positive*, or *surprised* in a given topic segment. We use established ASR CallHome train/dev/test partitions (80/20/20 conversations respectively) to train and evaluate our models. We report weighted F1 scores to evaluate the performance of our models.

All neural network models described in the following sections were implemented using PyTorch [21]. Inline with recent work on automatic emotion recognition, we focus on fairly generic lexical and acoustic input feature extraction methods (cf. [16, 22]), which are described in the following sections.

3.2. Acoustic Features

In the following experiments, we use the 88 dimensional eGeMAPS acoustic feature set [23] extracted using the OpenSMILE toolkit [24]. This includes a range of acoustic features that have been associated with affective states and has proven a strong baseline for automatic emotion recognition [23, 25]. These features are aggregate statistics (i.e. functionals) calculated over the frame-level pitch, energy, and spectral feature time series extracted for a given speech interval. In our experiments, we extract features for individual speakers (i.e. separate channels) over both turns and whole topic segments. We normalize the features to have zero mean and unit variance based on training set statistics. We found the eGeMAPs feature set provided as good or better performance than the much larger IS13 ComParE feature set (6373 features), cf. [16, 22], so we focus on the eGeMAPs related results for brevity.

3.3. Lexical Features

We build lexical representations over turns and topic segments using 300 dimensional GloVe word embeddings (Common Crawl, 840B tokens) [26]. We perform basic tokenization to map between the CallHome transcripts and word embeddings. We build up representations over turns and topic segments using Recurrent Neural Networks (RNNs) with Gated Recurrent Units (GRUs). In the following experiments, we use single layer GRU-RNNs, resulting in 100-dimensional lexical encodings over turns and/or topic segments. In each case, we take the RNN state for the last element in the sequence as the representation of that sequence. We also tried using the CNN based representations described in [22], but these generally performed worse than the GRU based representations, so we only report results for the latter here. We also found that allowing fine tuning of the word embeddings in our models did not improve performance, so only report results with fixed embeddings.

3.4. Hierarchical Models

We experiment with a number of ways to generate representations for topic segments using lexical and acoustic features. The first approach uses a single GRU-RNN to encode the speaker transcript for a topic segment (**topic.lex**), or similarly, eGeMAPS acoustic features over the entire topic segment (**topic.acoustic**), i.e. a *flat* approach. The second approach uses *hierarchical* RNNs: we obtain topic segment encodings as the output of a GRU-RNN over turn-level encodings. Acoustic turn-level encodings are derived by passing turn-level eGeMAPs features through a 300 dimensional feedforward layer (**turn.acoustic**), while lexical turn-level encodings are derived using a GRU-RNN over word embeddings in each turn (**turn.lex**).

Our goal is to detect the stance for a single target speaker in each topic segment. However, we would like to know if including information from the non-target speaker turns helps. We explore this using two variants. The first (default) variant uses only contributions of the target speaker, while in the second (**+listener**), we simply append a target speaker indicator variable to the turn-level encodings. We fuse lexical and acoustic representations by concatenating their encodings at the topic segment level. We also experiment with concatenating lexical and acoustic encodings at the turn level (**turn fusion**).

After obtaining a topic segment encoding, we apply a single 100 unit feedforward layer, which then feeds into a softmax output layer. In our experiments, we use tanh activations between layers. To prevent overfitting, we include 50% dropout layers between forward layers and early stopping based on validation set loss. We use the Adam optimizer to train our networks with a learning rate of 0.0001 and batch size of 128.

3.5. DialogueRNN Baseline

We also report results using a state-of-the-art emotion recognition architecture, DialogueRNN [16]. This model attempts to capture speaker interactions using a set of GRU cells to mediate updates to global and speaker specific dialogue states. At each turn, the current speaker's state is updated based on their current contribution, previous state and the global state. We use the *Bidirectional RNN+Attention* version of this model, which allows the emotion decoder to attend to all of the inferred emotion state representations of all turns in the dialogue, as well as the current emotion state. We also experimented with variants that include listener state updates at every time step (**active listening**), or keep the listener state static (**simple**).

DialogueRNN was designed to predict emotions at every speaker turn, so in training we broadcast the topic segment stance label to every turn in the segment. In evaluation, we take our stance prediction to be the majority vote over all the target speaker's turns in that segment. As input, we use 100 dimensional turn-level lexical and acoustic encodings (cf. Section 3.4). To make this more similar to the DialogueRNN setup, however, the turn-level encoders used to generate these features were trained to predict the topic stance label every turn, rather than only once per topic segment.

4. Results

Weighted F1 scores for experiments on the CallHome test set are shown in Table 1. Overall, including both lexical and acous-

Table 1: *Test set results: Weighted F1 scores. Overall best results are in **bold**.*

Model	Amused	Certain	Comfortable	Interested	Positive	Surprised
topic.acoustic	0.75	0.63	0.61	0.80	0.59	0.81
turn.acoustic	0.71	0.63	0.56	0.78	0.61	0.82
turn.acoustic +listener	0.65	0.45	0.54	0.78	0.48	0.80
topic.lex	0.65	0.63	0.57	0.78	0.63	0.80
turn.lex	0.65	0.64	0.54	0.78	0.64	0.80
turn.lex +listener	0.67	0.61	0.54	0.78	0.64	0.81
topic.lex, topic.acoustic	**0.78**	0.64	0.61	**0.80**	0.64	0.81
topic.lex, turn.acoustic	0.71	0.63	0.59	0.78	**0.67**	**0.82**
turn.lex, topic.acoustic	0.77	0.65	**0.61**	0.78	0.62	0.80
turn.lex, turn.acoustic	0.70	**0.65**	0.60	0.78	0.66	0.81
turn.lex, turn.acoustic (+listener)	0.67	0.60	0.54	0.78	0.65	0.80
turn.lex, turn.acoustic (turn fusion)	0.71	0.65	0.58	0.78	0.66	0.81
DialogueRNN (active listener)	0.69	0.63	0.54	0.78	0.64	0.81
DialogueRNN (simple)	0.67	0.61	0.54	0.78	0.63	0.80

tic features gives us the best performance. We get our best results for amusement, comfort, and interest using topic level acoustic features. This suggests that these stances are expressed over multi-turn segments. This is supported by the fact that topic level acoustic features provide the best unimodal performance for these stances. That is, it's not so much what people say but how they sound over multiple turns that tells you if someone is amused, comfortable or interested.

Inclusion of turn-specific information seems more important for detection of other stances. The hierarchical lexical models performed the best for detecting certainty and positivity, indicating that these stances are expressed using more overt lexical cues. Similarly, for surprise, we see a benefit from using turn level acoustic information. This indicates that expression of these stances is more localized within a topic segment, and using hierarchical encoders help expose this information. However, our best results often come from a combination of turn level and topic segment level inputs. This suggests that this task may benefit from structured hierarchical feature fusion strategies [25, 27] as well as hierarchical unimodal encodings.

Topic fusion strategies generally performed better than turn fusion. Nevertheless, we still found our turn fusion models generally perform better than DialogueRNN, which also fuses lexical and acoustic features at the turn level. This is somewhat surprising, as we would expect that modelling global/listener states to help stance detection. In fact, unlike [16], we generally see better results using the Active Listener variant even though we don't have non-speech features to drive listener state updates. However, this gives us less of an improvement than what we get from our best hierarchical models, even though they don't explicitly model interaction. Similarly, although it didn't generally improve performance for our hierarchical models, our simple approach of adding target speaker indicators to turn level encodings (+listener) performs at least as well as DialogueRNN with the simple listener. This indicates that further incorporation of long-term/hierarchical structure into the DialogueRNN architecture may be necessary to make use of its state tracking capabilities.

5. Discussion and Conclusion

In this paper, we investigated topic-oriented speaker stance in the CallHome corpus of conversational dialogues. Our experiments show that it is possible to automatically detect speaker stance on this challenging data set using fairly generic acoustic and lexical features. The results highlighted the importance of including both lexical and acoustic features in this task. Beyond this, we saw that some stances benefited more than others from additional hierarchical structure in our neural network models. This indicates that flexibility in modelling long- and short-term dialogue characteristics is important here, particularly with respect to acoustic features. The fact that our simple 'flat' models performed better than DialogueRNN for amusement and interest suggests that important long-term characteristics aren't being captured in its recurrent state update mechanisms. However, its ability to model speaker and listener states does appear useful for this task. So, we expect it would be beneficial to augment the DialogueRNN architecture to explicitly model topic-level information beyond it's current capabilities.

As noted above, DialogueRNN was designed to predict turn-level emotions. While closely related, stance detection presents a different view of affect in the data. In particular, our results suggest that positive affect can be expressed differently depending on what stance is most salient at that point in the dialogue. Most current emotion recognition work focuses on the 'Big Six' emotion categories [28, 7], which only includes one clearly positive category: happiness. So, analyses based on these categories are likely to lump these different expressions of positivity together. However, our stances are more closely aligned with multi-dimensional views of emotion [29]. While they're not reducible to these dimensions, we'd expect our notion of positive topic-oriented stance to correlate with valence ratings, surprise to expectancy, interest to arousal, and comfort to power. As such, future work will look at whether pre-training models with emotion annotated data helps stance detection. We also plan to look at how emotion predictions relate to our stance annotations, as this may help shed light on the types of affective states are prominent in everyday conversational dialogue. We also plan to perform more detailed lexical/phonetic analysis of stances cues to help guide speech generation/synthesis.

Beyond the relationship between stance and emotions, we also need to consider the topic part of topic-oriented stance. In this work, we relied on manually identified topic segments. However, to actually detect topic-oriented stance in conversations, we also need to detect topics. Previous work on topic detection in speech has generally focused on linear segmentation and labeling of broadcast news [30, 31]. However, topic boundaries in conversational dialogue are less clearly/linearly defined [32, 33, 34]. Thus, our next steps will investigate methods for joint topic and stance detection, and whether including topic descriptions as inputs can help detect stance.

6. References

[1] J. W. Du Bois, "The stance triangle," *Stancetaking in discourse: Subjectivity, evaluation, interaction*, vol. 164, pp. 139–182, 2007.

[2] K. Forbes-Riley and D. J. Litman, "Adapting to student uncertainty improves tutoring dialogues." in *AIED*, 2009, pp. 33–40.

[3] M. A. Walker, P. Anand, R. Abbott, J. E. F. Tree, C. Martell, and J. King, "That is your evidence?: Classifying stance in online political debate," *Decision Support Systems*, vol. 53, no. 4, pp. 719–729, 2012.

[4] S. M. Mohammad, P. Sobhani, and S. Kiritchenko, "Stance and sentiment in tweets," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 3, p. 26, 2017.

[5] P. Sobhani, D. Inkpen, and X. Zhu, "A dataset for multi-target stance detection," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, vol. 2, 2017, pp. 551–557.

[6] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[7] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[8] J. Williams, S. Kleinegesse, R. Comanescu, and O. Radu, "Recognizing emotions in video using multimodal dnn feature fusion," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, 2018, pp. 11–19.

[9] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 2011, pp. 169–176.

[10] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[11] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intelligent Systems*, vol. 33, no. 6, pp. 17–25, 2018.

[12] N. G. Ward, J. C. Carlson, and O. Fuentes, "Inferring stance in news broadcasts from prosodic-feature configurations," *Computer Speech & Language*, vol. 50, pp. 85–104, 2018.

[13] V. Freeman, J. Chan, G.-A. Levow, R. Wright, M. Ostendorf, and V. Zayats, "Manipulating stance and involvement using collaborative tasks: An exploratory comparison," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] G.-A. Levow, V. Freeman, A. Hrynkevich, M. Ostendorf, R. Wright, J. Chan, Y. Luan, and T. Tran, "Recognition of stance strength and polarity in spontaneous speech," in *2014 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2014, pp. 236–241.

[15] V. Freeman, "The phonetics of stance-taking," Ph.D. dissertation, University of Washington, 2015.

[16] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, E. Cambria, and G. Alexander, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations," in *AAAI 2019*.

[17] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[18] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[19] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii, "Brat: a web-based tool for nlp-assisted text annotation," in *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 102–107.

[20] K. Krippendorff, "Reliability in content analysis: Some common misconceptions and recommendations," *Human communication research*, vol. 30, no. 3, pp. 411–433, 2004.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[22] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L. Morency, and R. Zimmermann, "Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos," in *Proceedings of NAACL 2018*, 2018.

[23] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[25] L. Tian, J. Moore, and C. Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 565–572.

[26] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[27] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge-Based Systems*, vol. 161, pp. 124–133, 2018.

[28] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2018, pp. 2236–2246.

[29] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[30] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating Prosodic and Lexical Cues for Automatic Topic Segmentation," *Computational Linguistics*, vol. 27, no. 1, pp. 31–57, Mar. 2001.

[31] E. Tsunoo, O. Klejch, P. Bell, and S. Renals, "Hierarchical recurrent neural network for story segmentation using fusion of lexical and acoustic features," in *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*. IEEE, 2017, pp. 525–532.

[32] B. Grosz and C. Sidner, "Attention, intentions, and the structure of discourse," *Computational linguistics*, vol. 12, no. 3, pp. 175–204, 1986.

[33] R. J. Passonneau and D. J. Litman, "Discourse Segmentation by Human and Automated Means," *Computational Linguistics*, vol. 23, no. 1, pp. 103–139, Mar. 1997.

[34] J. Niekrasz and J. Moore, "Participant subjectivity and involvement as a basis for discourse segmentation," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2009, pp. 54–61.