

Computer Assisted Annotation of Tension Development in TED Talks through Crowdsourcing

Seungwon Yoon Wonsuk Yang Jong C. Park[†]

School of Computing

Korea Advanced Institute of Science and Technology

{swyoon, derrick0511, park}@nlp.kaist.ac.kr

Abstract

We propose a method of machine-assisted annotation for the identification of tension development, annotating whether the tension is increasing, decreasing, or staying unchanged. We use a neural network based prediction model, whose predicted results are given to the annotators as initial values for the options that they are asked to choose. By presenting such initial values to the annotators, the annotation task becomes an evaluation task where the annotators inspect whether or not the predicted results are correct. To demonstrate the effectiveness of our method, we performed the annotation task in both in-house and crowdsourced environments. For the crowdsourced environment, we compared the annotation results with and without our method of machine-assisted annotation. We find that the results with our method showed a higher agreement to the gold standard than those without, though our method had little effect at reducing the time for annotation. Our codes for the experiment are made publicly available¹.

1 Introduction

Recently, researchers for natural language processing are paying more attention to crowdsourcing for its effectiveness in linguistic annotations. The recent development in crowdsourcing platforms such as Amazon Mechanical Turk (AMT) has much reduced the time and effort required for an annotation project. Many researchers proposed methods to assist the workers in the crowdsourced annotation (Yuen et al. (2011); Poesio et al. (2013); Guillaume et al. (2016); Madge et al. (2019); Yang et al. (2019)). In particular, Guillaume et al. (2016) designed a game-based platform for the annotation of dependency relations in

French text, with the prediction model embedded in their platform. Yang et al. (2019) proposed to predict the difficulty of an annotation unit in order to allocate relatively easy units to crowdsourcing workers and the rest to expert annotators.

In this paper, we present a machine-assisting method for effective annotation of tension development. Tension is a means to keep the attention of the reader or audience, studied mainly in the field of storytelling (Zillmann (1980); Klimmt et al. (2009); Niehaus and Young (2014)). Tension also plays a critical role in discourse development (Lehne and Koelsch, 2015). We annotate the tension development, whether the tension is increasing, decreasing, or staying unchanged, in the TED Talks. We also introduce a Self-Assessment Manikin (SAM), which is an intuitive diagram that helps understand the annotation guidelines for tension annotation. Our method uses a prediction model for tension development, and provides the annotators with model predicted results as initial values. The predictions are based on the audio, the subtitle of the given video clip and the previous annotation results by an annotator.

We validate our method through an experiment on crowdsourced annotations. The annotations with our method show a higher agreement to the gold standard, which we instructed manually by annotating independently from the crowdsourced annotations, than those without our method. However, contrary to our initial expectation that our method will also reduce the annotation time, we find that it hardly reduced the time.

The contributions of this paper are as follows. (1) We proposed a new annotation scheme using the Self-Assessment Manikin (SAM) to annotate the tension development on multimodal data. (2) To the best of our knowledge, our method is the first in utilizing a prediction model to assist the annotation of tension development. We show experimen-

[†]Corresponding author

¹<https://github.com/nlpcl-lab/ted-talks-annotation>

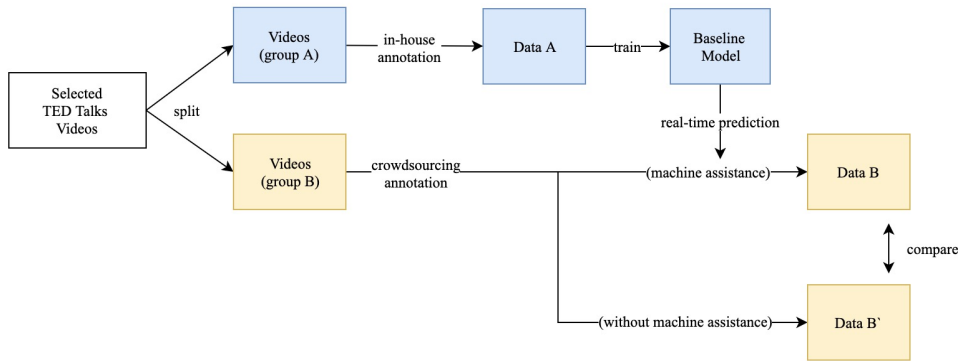


Figure 1: Overview of the annotation process

tally that our method is effective at gathering high-quality data and provide a detailed analysis of the annotation results. (3) We make the related data and the code publicly available.

2 Related work

2.1 Computer-Assisted Annotation

Ringger et al. (2008) suggested a machine-assisted method for part-of-speech (POS) tagging. They provided model predicted results to the annotators so that the annotators may focus only on incorrect predictions. There has been a line of researches for effective visualization and an improvement on the user-interface that can help a linguistic annotation process (Stenetorp et al. (2012); Yimam et al. (2013)). Guillaume et al. (2016) provided a game-based platform for the annotation of dependency relations in French text and used a prediction model as a part of the platform in the training phase for the annotators before the main data gathering. For the selection of the target data to annotate, active learning has been employed to selectively collect only the training data on which the model does not perform well in order to maximize the performance of the model with a dataset that is as small as possible (Wang et al. (2017); Duong et al. (2018)). Schulz et al. (2019) showed that the provision of the automatically generated annotation results can accelerate the annotation process and enhance the annotation quality, without incurring a significant bias.

For visual object detection, Yao et al. (2012) presented an annotation platform that contains a prediction model for the location of the given object. In their platform, the model presents the predicted location to the annotators, and the annotators modified the location if it is incorrect. They

also predicted the time that the annotator may take for the modification and presented the annotation unit to the annotators with the shortest expected time to minimize the total cost of their annotation project. Su et al. (2012) presented a quantification test that can identify the annotators who do not fully understand the annotation guidelines. They also presented a rule-based feedback system that can warn untrained annotators before continuing the annotation.

2.2 Emotion, suspense, and tension

Tension is a psychological concept that is related to emotion and suspense. Tension has been studied along with suspense for the literature, movies, and games (Brewer and Lichtenstein (1982); Zillmann (1980); Klimmt et al. (2009)). Lehne and Koelsch (2015) proposed a general psychological model for tension without any further restriction on its domain, defining the magnitude of tension as the interval between positive and negative expectations of the outcome.

In the field of computer science, there has been a line of researches modeling the mental state of the reader to create an intense story (Niehaus and Young (2014); O’Neill and Riedl (2014)). Li et al. (2018) designed a scheme for story structures considering dramatic tension changes and the narrative structure suggested by Helm and MacNeish (1967) and annotated the story structure for short stories and personal anecdotes. For the analysis of emotion, Cowie and Sawey (2011) annotated on the intensity of laughter and the degree of positive emotion in the videos of babies. Metallinou and Narayanan (2013) annotated on activation, valence, and dominance with an assumption that the three attributes represent the state of emotion in video. Antony et al. (2014) annotated changes in

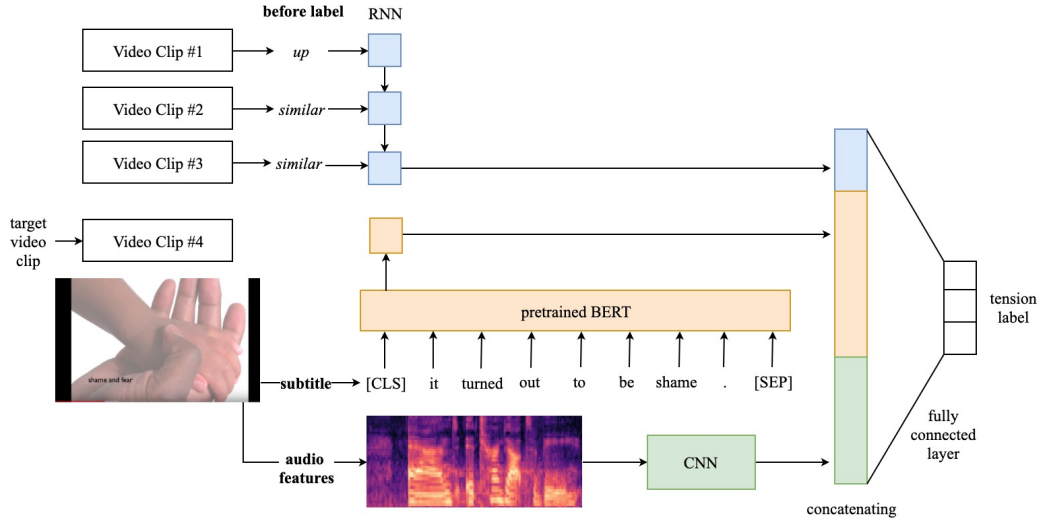


Figure 2: Model architecture

arousal and valence with heart rate, electrodermal activity, and respiration rate. The multi-modal data collection enables a more flexible analysis of the environmental interactions.

3 Data

We used the TED Talks as a dataset to track the tension development. TED Talks are a conference that presents ideas on various topics in a few minutes, and the video part has been used for emotional analysis and assessment of engagement exploiting the highly reliable English subtitles precisely synchronized to the video (Neumann and Vu (2019); Haider et al. (2017))). For the annotation of tension development, we have chosen to use TED Talks with two specific reasons: (1) Due to the nature of public lectures, many utterances raise the tension to keep the attention of the audience. (2) The applause or laughter of the audience, which may be highly related to tension development, is also recorded in the video.

In the archives of TED Talks², we randomly selected 20 videos whose running time is in the range of 10-20 minutes. For each of the 20 videos, we divided it into a set of small video clips, where the division was based on the subtitles so that a clip corresponds to a sentence. The English subtitles were split into sentences. We obtained a dataset containing 3,597 video clips with a total duration of 301 minutes. Each sentence that corresponds to a video clip consists of 14 words on

²Videos and subtitles at <http://www.ted.com> are publicly available under Creative Commons license, Attribution–Non Commercial–No Derivatives.

average.

4 Method

Our method uses a neural network based prediction model, and provides the predicted results to the annotators as the initial values for the options that the annotator is asked to fill out. By this, the annotation task, originally to choose the correct label for a given video clip, is transformed into an evaluation task, judging whether or not the predicted result by the model is correct.

Figure 2 shows the architecture of our model. The model predicts the label for each video clip sequentially, and utilizes three features: subtitles, audio, and the formerly chosen labels for the previous video clips. The audio of a video clip was encoded into a vector using CNN. We used pyAudioAnalysis software (Giannakopoulos, 2015) to extract 34 features such as MFCC at the rate of 30 frames/sec, and the features were passed to the CNN. The CNN consists of three 1D convolutional layers. 1D max-pooling with ReLU activation function is performed after each convolutional layer. The lecture’s subtitles were encoded into a vector using a pre-trained uncased BERT-base model (Devlin et al., 2019). The previously chosen k labels were encoded into a vector using an RNN. The three vectors for the three features were concatenated into a vector, passed afterwards to the output layer, or the fully connected layer.

type	#videos	#video clips
in-house	10 (group A)	1,736
crowdsourced	10 (group B)	1,861
all	20	3,597

Table 1: Statistics of the data

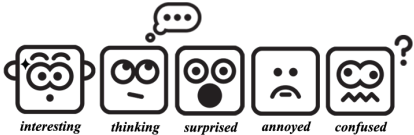

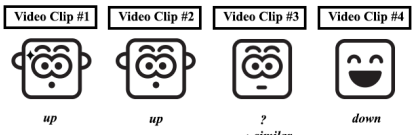
label (score)	guidelines
up (+1)	<p>Watch the video clip and select <i>up</i> if your feeling matches one of the pictures below.</p>  <p>interesting: I'm interested, want to learn more and know what's next. thinking: I'm thinking about the content of the lecture (e.g., when the speaker asks a question). surprised: I'm surprised at seeing something I didn't expect. annoyed: I'm uncomfortable or feeling that the content is unpleasant or difficult to agree with. confused: I'm confused because it is different from what I originally knew or it is difficult to understand.</p>
down (-1)	<p>Watch the video clip and select <i>down</i> if your feeling matches one of the pictures below.</p>  <p>relieved: I am comfortable again, due to the removal of any previous anxiety or doubt. funny: I find the speaker's joke(s) or content to be amusing. boring: I am not interested in the repetition of similar and/or uninteresting content.</p>
similar (0)	<p>Watch the video clip and select <i>similar</i> when your status is neither <i>up</i> nor <i>down</i>. If you are uncertain about your feeling, as shown in the third video clip of the picture below, select <i>similar</i>.</p> 

Table 2: Annotation guidelines for the change in tension

5 Annotation

5.1 Overview

Figure 1 gives an overview of our annotation of tension development. First, as shown in Table 1, 10 TED Talks videos were divided into group A and group B. In-house annotation was performed on group A and the results, which we call data A, were used for training the prediction model. Then, group B was annotated through Amazon Mechanical Turk (AMT), a crowdsourcing platform. For group B, the crowdsourced annotation was conducted in two phases. First, every video in group B was annotated via AMT *using* our method (data B). Second, independently of the first, every video in group B was annotated via AMT, *not using* our method (data B').

For a video, the annotators watched the video clips in their original order, and annotated on each clip with one of the three labels, *up*, *down*, and *similar*. *Up* indicates that the tension is increasing, and *down* indicates that it is decreasing. *Similar* indicates that the tension is not changing. As it is disruptive for the annotator to iterate the clicking on the video for playing and pausing, we made an annotation tool to prevent such disruption (Figure 3).

Due to the copyright issue, we could not post the TED Talks video directly online. Instead, we provided the annotators, or crowdsourcing workers, with the videos at TED's official Youtube channel³ via an embedded player, controlled by the APIs provided by the Youtube player. If the annotator enters a shortcut key to move to the next video clip or presses the play button of the video clip, the video clip is played. After the video clip meets the end (of the clip), an input window for annotation is displayed. Then, the annotator can perform the annotation on the clip, and proceed to the next clip. We also provided the subtitles explicitly to the annotators.

5.2 Annotation Scheme

The tension development within each video clip was annotated with one of the three values (*up*, *down*, or *similar*). We defined each of the three labels based on the specific circumstances in Table 2. Five circumstances, which are interesting,

³<https://www.youtube.com/user/TEDtalksDirector>

⁴source of the video: <https://www.youtube.com/watch?v=iCvmsMz1F7o>

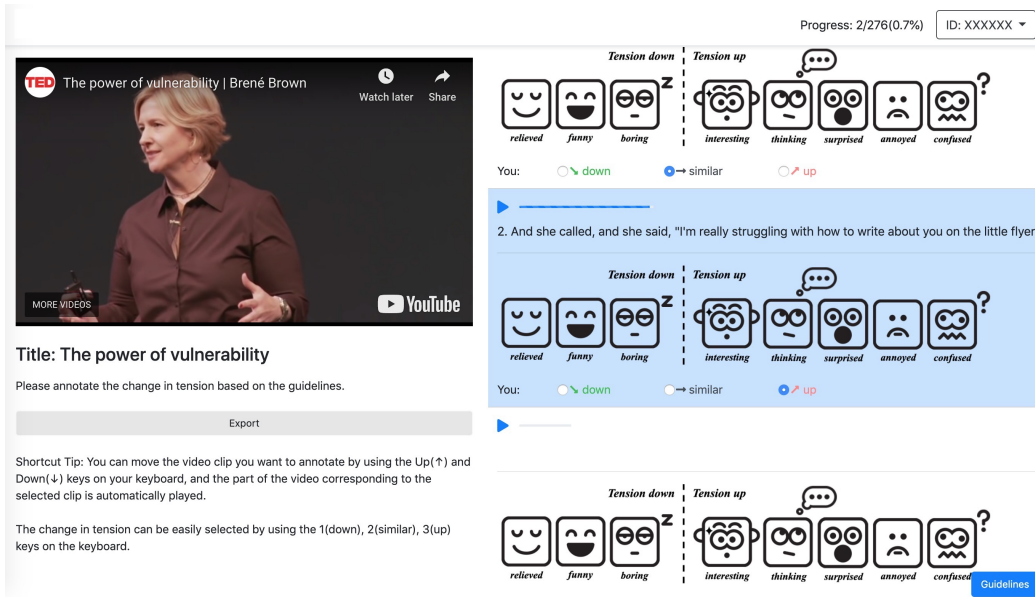


Figure 3: Interface of the annotation tool⁴

thinking, surprised, annoyed, and confused, correspond to *up*. If a video clip can be described as one of the five circumstances, we defined the video clip to have the label of *up*. In a similar way, three circumstances, or relieved, funny, and boring, correspond to the label of *down*. If a video clip is judged to be neither *up* nor *down*, we defined it as having the label of *similar*. It should be noted that the definition of the labels is designed specifically for the domain of public lectures. For example, ridiculing someone in everyday life may increase the tension. Still, in lectures, it is often intended to help the audience to feel relaxed and help them to feel comfortable listening (Meyer, 2000). Therefore, we set it as a circumstance for *down*.

To help the annotators to intuitively follow up the annotation guidelines, and for the cases where the annotators forget the details of the guidelines (of the specification of the circumstances), we provided Self-Assessment Manikins (SAMs) to the annotators as shown in Table 2. Providing SAMs to annotators has been acknowledged to be an effective method for an emotion-related annotation task (Bradley and Lang (1994); Yadati et al. (2013); Boccignone et al. (2017)).

5.3 Annotation Procedure

5.3.1 In-house Annotation

The in-house annotation method was used to annotate 1,736 video clips (group A). A total of five annotators participated, and three annotators anno-

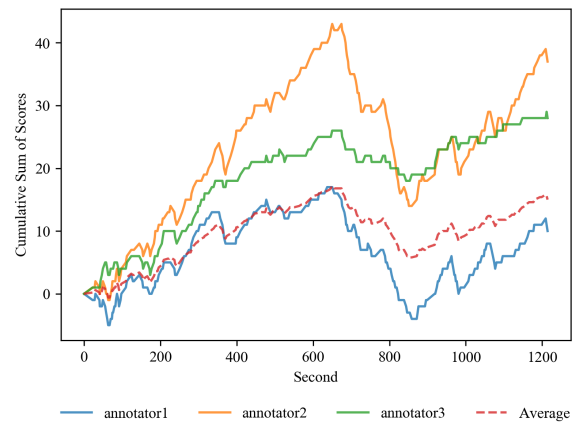


Figure 4: Example of annotations by three annotators

tated the video clips for each video. 5,208 annotation values were obtained for 10 videos containing 1,736 video clips. The distribution of *down*, *similar*, and *up* labels was 749 (14.4%), 3,218 (61.8%), and 1,239 (23.8%), respectively.

Figure 4 illustrates an example of the cumulative sum of scores annotated by three annotators for the same video. The chosen values were slightly different among the annotators (Krippendorff's α : 0.298), but the tendency to exceed or fall short of the cumulative sum of scores was similar (mean correlation: 0.73). Since each annotator has a different personal scale by which to rate emotion, Pearson's correlation and Cronbach's α , which are indicators that focus on trends when evaluating the agreement of annotation,

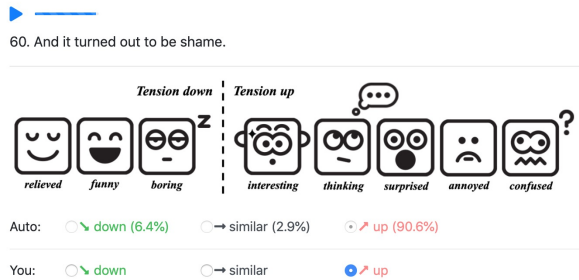


Figure 5: Interface showing predicted values in machine-assisted annotation

were used (McKeown et al. (2011); Metallinou and Narayanan (2013)). For in-house annotations, we obtained the agreements as shown in Table 5. Pearson’s correlation and Cronbach’s α were measured as the cumulative sum of the scores.

type	down	similar	up	sum
train	153	819	243	1,215
test	69	342	110	521
all	222	1,161	353	1,736

Table 3: Statistics of the data for training the model

5.3.2 Crowdsourcing Annotation

Of the data collected via in-house annotations to the Group A videos, 70% were used as the training set and 30% were used as the test set to train and evaluate the model (Table 3). When setting the ground truth from data annotated by three people in the same video clip, we decided to use majority voting among *down*, *similar*, and *up* labels. If each label was selected once, the label *similar* was set as the ground truth.

When annotating with crowdsourcing, the videos in group B were annotated with and without machine assistance by three annotators each (Figure 5). Video clips annotated without machine assistance were annotated using the same interface as used for the in-house annotation. During machine-assisted annotation, predicted values by the model are presented along with the probability, and the label with the highest probability was given to the annotator as the default value. The trained model provided predicted values in real-time using the subtitles, sound of the video clips and the tension values that the user annotated in the previous five video clips. Annotators were instructed to refer to the automatic prediction value:

ground truth	down	similar	up
down	41	23	5
similar	17	282	44
up	1	63	45

Figure 6: Confusion matrix of the prediction model on the test set

“Please note that the value of the predicted tension is automatically given as the default value. If your judgment is different, change the value according to your judgment. If the default value matches your judgment, you may move on to the next video clip.”

We used the Amazon Mechanical Turk (AMT) service for crowdsourcing, providing workers with annotation guidelines and the URL for the web-based annotation tool. Each worker was allowed to participate in annotating several different videos. Workers with the number of HITs approved > 50 and HIT approval rate $> 95\%$ were allowed to join. There were a total of 47 annotators.

feature	Precision	Recall	F1
audio	0.54	0.50	0.52
text	0.61	0.60	0.60
before label (k=5)	0.43	0.49	0.45
audio + text	0.65	0.61	0.63
+ before label (k=5)			

Table 4: Comparison of the performance on the test set according to the features used

5.3.3 Analysis of Annotations

Figure 6 shows the confusion matrix of the prediction model in the test set. The performance (F1 score) for the down label (0.64) was higher than that for up (0.44). Table 4 compares the performance according to the features used. The performance was lowest when the tension labels of the previous video clip were used as a feature. It was highest when they used three types of features together.

type	video group	#annotator for each video clip	agreement			mean selection time (seconds)
			mean Pearson's correlation	mean Cronbach's α	Krippendorff's α	
in-house	group A	3	0.645	0.855	0.283	2.02
crowdsourced	machine assistance	group B	0.817	0.817	0.387	2.61
	no machine assistance	group B	0.636	0.469	0.134	2.69

Table 5: Statistics for agreement, time of annotation results

As the result of the annotation, 11,166 annotation values were obtained for 10 videos with 1,861 video clips (group B). For machine-assisted annotations, the distribution of *down*, *similar* and *up* was 895 (16.0%), 2,862 (51.3%), and 1,826 (32.7%), respectively. For unassisted annotations from the machine, the distribution was 977 (17.5%), 2,372 (42.4%), and 2,232 (39.9%). Table 5 shows the agreement among the annotation results. In-house annotations were all higher in all the three metric than the crowdsourced annotations without machine-assistance. In the control group, machine-assisted annotations showed higher levels of agreement than non-assisted annotations.

We analyzed whether the improvement of agreement rate was a negative effect from the bias resulting from the predicted labels. For analysis, gold labels were compared to annotations. Gold labels were set by the annotations of one of the authors with no machine assistance in 4 videos selected in group B. Figure 7 shows an example of such gold labels, machine-assisted annotations and the annotations of the control group for the cumulative sum of the tension score. Comparing the mean correlation for the 4 videos, the mean correlation of the machine-assisted annotations was 0.861, higher than the control group's mean correlation of 0.466. The annotation values were more in line with the trend among gold labels with machine-assistance.

The mean correlation between machine predictions itself and gold labels was 0.867. This means that machine-assisted annotators can achieve results closer to gold than the control group if they accept all the predicted values. However, machine-assisted annotators changed 26.5% of the labels presented as default values through the model (Figure 8). The change ratio of prediction values for each of *down*, *similar* and *up* is 17.7%, 28.8% and 24.3%, respectively. This produced a difference between machine predictions and machine-assisted annotations, as illustrated in Figure 7. The

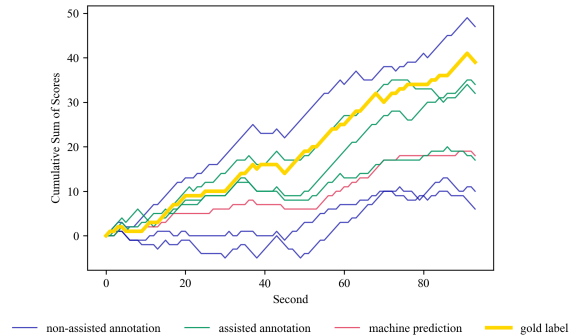


Figure 7: Example of annotations with gold label

average of the probabilities (as shown in Figure 5) presented with labels set as default values by the prediction model was 90.4%. When the user changed the default value, the average of the probabilities was 87.0%. When the user did not change the default value, the average was 91.6%.

The selection times in Table 5 represent the amount of time it takes to select the tension label from the time the video clip is played to the end. For machine-assisted annotations, if the default value is not changed by the annotator, the time between the end of the current video clip and the start of the next video clip was considered as the selection time. When receiving machine assistance, the annotation time was expected to be reduced because the input process of selecting labels would disappear if the model prediction values and the annotator's judgments were the same. However, there was no significant difference compared to the control group.

6 Conclusion

In this paper, we introduced a method for machine-assisted annotation of tension development. Our method utilizes a prediction model to provide the predicted result to the annotators so that the annotation task is turned into an evaluation task of inspecting whether or not the prediction by the model is correct. We find that our method enhances the agreement of the crowdsourced anno-

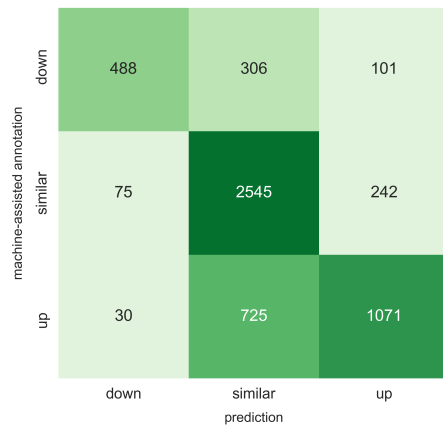


Figure 8: Confusion matrix of the prediction model on the group B videos

tations to the gold standard annotation in a small trial of 3 annotators. We also find that our method does not particularly affect the time taken for the annotation.

We proposed a new annotation scheme using the Self-Assessment Manikin (SAM) to annotate the tension development. By converting the annotation task into a verification task via machine assistance, the results become consequently more aligned with the gold standard compared with the control group.

Acknowledgments

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582-002, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection). We thank the anonymous reviewers for the much helpful feedback.

References

- J Antony, K Sharma, C Castellini, Egon L van den Broek, and C Borst. 2014. Continuous affect state annotation using a joystick-based user interface. In *Proceedings of Measuring Behavior*, pages 268–271.
- Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. 2017. Amhuse: a multimodal dataset for humour sensing. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 438–445. ACM.
- Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the se-

mantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59.

William F Brewer and Edward H Lichtenstein. 1982. Stories are to entertain: A structural-affect theory of stories. *Journal of pragmatics*, 6(5-6):473–486.

Roddy Cowie and Martin Sawey. 2011. Gtrace-general trace program from queen’s, belfast. <https://sites.google.com/site/roddycowie/work-resources>.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 43–48.

Theodoros Giannakopoulos. 2015. pyaudioanalysis: An open-source python library for audio signal analysis. *PloS one*, 10(12):e0144610.

Bruno Guillaume, Karën Fort, and Nicolas Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Fasih Haider, Fahim A Salim, Saturnino Luz, Carl Vogel, Owen Conlan, and Nick Campbell. 2017. Visual, laughter, applause and spoken expression features for predicting engagement within TED talks. *Feedback*, 10:20.

June Helm and June Helm MacNeish. 1967. *Essays on the verbal and visual arts*. University of Washington Press.

Christoph Klimmt, Albert Rizzo, Peter Vorderer, Jan Koch, and Till Fischer. 2009. Experimental evidence for suspense as determinant of video game enjoyment. *CyberPsychology & Behavior*, 12(1):29–31.

Moritz Lehne and Stefan Koelsch. 2015. Toward a general psychological model of tension and suspense. *Frontiers in Psychology*, 6:79.

Boyang Li, Beth Cardier, Tong Wang, and Florian Metzger. 2018. Annotating high-level structures of short stories and personal anecdotes. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

- Chris Madge, Juntao Yu, Jon Chamberlain, Udo Kruschwitz, Silviu Paun, and Massimo Poesio. 2019. Crowdsourcing and aggregating nested markable annotations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 797–807.
- Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. 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.
- Angeliki Metallinou and Shrikanth Narayanan. 2013. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.
- John C Meyer. 2000. Humor as a double-edged sword: Four functions of humor in communication. *Communication theory*, 10(3):310–331.
- Michael Neumann and Ngoc Thang Vu. 2019. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394. IEEE.
- James Niehaus and R Michael Young. 2014. Cognitive models of discourse comprehension for narrative generation. *Literary and linguistic computing*, 29(4):561–582.
- Brian O’Neill and Mark Riedl. 2014. Dramatis: A computational model of suspense. In *28th AAAI Conference on Artificial Intelligence*, pages 944–950.
- Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, and Luca Ducceschi. 2013. Phrase detectives: Utilizing collective intelligence for internet-scale language resource creation. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(1):3.
- Eric K Ringger, Marc Carmen, Robbie Haertel, Kevin D Seppi, Deryle Lonsdale, Peter McClanahan, James L Carroll, and Noel Ellison. 2008. Assessing the costs of machine-assisted corpus annotation through a user study. In *Proceedings of LREC*, volume 8, pages 3318–3324.
- Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych. 2019. [Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772, Florence, Italy. ACL.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. 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*, pages 102–107. ACL.
- Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. In *Workshops at the 26th AAAI Conference on Artificial Intelligence*.
- Chenguang Wang, Laura Chiticariu, and Yunyao Li. 2017. Active learning for black-box semantic role labeling with neural factors. In *Proceedings of IJCAI*, pages 2908–2914.
- Karthik Yadati, Harish Katti, and Mohan Kankanalli. 2013. Cavva: Computational affective video-in-video advertising. *IEEE Transactions on Multimedia*, 16(1):15–23.
- Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C. Wallace, and Ani Nenkova. 2019. [Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1471–1480, Minneapolis, Minnesota. ACL.
- Angela Yao, Juergen Gall, Christian Leistner, and Luc Van Gool. 2012. Interactive object detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3242–3249. IEEE.
- Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6.
- Man-Ching Yuen, Irwin King, and Kwong-Sak Leung. 2011. A survey of crowdsourcing systems. In *Proceedings of IEEE Third International Conference on Privacy, Security, Risk and Trust and IEEE Third International Conference on Social Computing*, pages 766–773. IEEE.
- Dolf Zillmann. 1980. Anatomy of suspense. In *The entertainment functions of television*, pages 133–163. Hillsdale, NJ: Erlbaum.