Sentiment Analysis for Emotional Speech Synthesis in a News Dialogue System

Hiroaki Takatsu[1] **Ryota Ando**[2] **Yoichi Matsuyama**[1] **Tetsunori Kobayashi**[1]
[1]Waseda University, Tokyo, Japan
[2]Naigai Pressclipping Bureau,Ltd., Tokyo, Japan
takatsu@pcl.cs.waseda.ac.jp, ando@naigaipc.co.jp,
matsuyama@pcl.cs.waseda.ac.jp, koba@waseda.jp

Abstract

As smart speakers and conversational robots become ubiquitous, the demand for expressive speech synthesis has increased. In this paper, to control the emotional parameters of the speech synthesis according to certain dialogue contents, we construct a news dataset with emotion labels (“positive,” “negative,” or “neutral”) annotated for each sentence. We then propose a method to identify emotion labels using a model combining BERT and BiLSTM-CRF, and evaluate its effectiveness using the constructed dataset. The results showed that the classification model performance can be efficiently improved by preferentially annotating news articles with low confidence in the human-in-the-loop machine learning framework.

1 Introduction

As smart speakers and conversational robots become ubiquitous, the demand for expressive speech synthesis has increased. Speech synthesis is a technology that converts input text into a speech signal that represents its contents. In recent years, research has not only improved the quality so that a synthesized voice is similar to a real one, but also has diversified expressions (Govind and Prasanna, 2013; Kaur and Singh, 2015; Skerry-Ryan et al., 2018). Emotional speech synthesis is a technique for diversifying the expression of speech synthesis (Qin et al., 2006; Schröder, 2001; Yang et al., 2018). It specifies both emotional parameters and the text input so that the speech reflects the designated emotion (Charfuelan and Steiner, 2013; Inoue et al., 2017; Iwata and Kobayashi, 2011; Nose and Kobayashi, 2013).

In conversational robots and non-task-oriented dialogue systems, the emotions of speech synthesis are controlled not only by the contents of the text spoken by the system but also by the personality of the robot or environmental information (Bennewitz et al., 2007; Chiba et al., 2018; Robbel et al., 2009). On the other hand, in audiobooks and news reading systems, the emotions when the system speaks are often determined by the content of the text. In such applications, it is desirable that the emotional parameters used in the speech synthesis system be automatically estimated from the input text (Bellegarda, 2011; Jauk et al., 2018; Shaikh et al., 2009; Sudhakar and Bensraj, 2014; Trilla and Alías, 2013; Vanmassenhove et al., 2016).

When news is transmitted by a synthesized voice, it is beneficial for listeners that news with positive content is transmitted with voices that are synthesized with positive emotion, whereas news with negative content is transmitted with voices that are synthesized with negative emotion (Pitrelli et al., 2006). In our spoken dialogue system that delivers news (Takatsu et al., 2018), it is important to speak clearly with emotion according to the content of the news to improve the users’ understanding. Table 1 shows an example of a news conversation. S_2 is “positive” because it reports that the subject, Nishikori, won the game, while S_4 is “negative” because it reports that the subject, Andy Murray, lost in the game. S_1 and S_3 are “neutral” because they contain neither positive nor negative content. To realize emotional speech synthesis according to such news content, herein we construct a dataset for machine learning by annotating emotion labels (“positive,” “negative,” or “neutral”) for each sentence in news articles and propose

Table 1: Example of emotions of each system utterance in conversational news delivery (S is system utterance, U is user utterance)

Utterance	Emotion
S_1 : ブリスベンでテニスの試合が行われてるんだけど A tennis match is taking place in Brisbane.	Neutral
S_2 : 錦織がデニス・クドラに勝ったんだって Nishikori beat Denis Kudla.	Positive
U_1 : さすが！ それで、次は誰と対戦するの？ That's great! So, who will he face in the next round?	
S_3 : 準々決勝ではグリゴール・ディミトロフと対戦するそうだよ He will face Grigor Dimitrov in the quarterfinals.	Neutral
U_2 : そうなんだ。アンディ・マリーはどうだったの？ I see. What was the Andy Murray's match result?	
S_4 : 元世界 1 位のアンディ・マリーは 2 回戦で負けちゃったみたいだね Former world number one, Andy Murray, lost in the second round.	Negative
U_3 : それは残念だね Oh, that's too bad.	

a method to identify them using a model that combines BERT (Devlin et al., 2019) and BiLSTM-CRF (Lample et al., 2016). Furthermore, we show that the model performance can be efficiently improved by repeating active learning in the human-in-the-loop machine learning framework (Munro, 2019).

The structure of this paper is as follows. Section 2 discusses related work. Section 3 overviews the annotation method of emotion labels and the statistics of the constructed dataset. Section 4 describes the proposed model. Section 5 shows the performance results of the proposed model. Section 6 reports the effects of active learning. Section 7 provides conclusions and future work.

2 Related work

Sentiment analysis is a technique to analyze emotions contained in texts (Hu et al., 2018; Zhang et al., 2018). Most sentiment analysis studies target texts containing subjective opinions such as Twitter tweets and review documents (Fan et al., 2018; Islam et al., 2019; Yin et al., 2019; Zhang and Zhang, 2019). Many of these studies classify emotions from the perspective of writers. On the other hand, studies on news articles that mainly refer to objective events classify emotions from the readers' perspective. For example, Lin et al. classified readers' emotions about news into happy, angry, sad, surprised, heartwarming, awesome, bored, and useful (Lin et al., 2007; Lin et al., 2008). They assumed that the most common emotion chosen by users who read the news article to be the correct answer and identified the emotions of news articles using SVM (Cortes and Vapnik, 1995). Li et al. classified readers' emotions into touching, empathy, boredom, anger, amusement, sadness, surprise, and warmness (Li et al., 2016). They formulated the sentiment analysis problem as a multi-label classification problem and proposed a topic model to estimate the weight of different documents for each emotion (Li et al., 2016). Ciptadi et al. classified readers' emotions into proud，angry，sad，happy，afraid，amused，inspired，and surprised (Ciptadi and Girsang, 2019). They showed that the classification performance of a Naive Bayes classifier and logistic regression was improved by applying the SMOTE (Chawla et al., 2002) oversampling method to alleviate the problem of imbalanced data.

In addition to studies classifying entire news articles (Ciptadi and Girsang, 2019; Ling et al., 2017; Li et al., 2016; Lin et al., 2007; Lin et al., 2008; Liu et al., 2013; Wang and Liu, 2017), some research classified headlines of news articles (Kirange and Deshmukh, 2012; Strapparava and Mihalcea, 2007), while others classified sentences of news articles (Bhowmick et al., 2009; Bhowmick et al., 2010; Das and Bandyopadhyay, 2009; Li et al., 2015; Patil and Chaudhari, 2012). In studies on sentence classification, for example, Bhowmick et al. classified readers' emotions into anger, disgust, fear, happiness, sadness, and surprise (Bhowmick et al., 2010). They asked multiple people to annotate the sentences of news articles. They confirmed that the agreement rate could be improved by eliminating surprise as well as integrating anger and disgust. In addition, they evaluated the multi-label classification performance by ADTboost.MH (Comité et al., 2003). Similarly, eliminating surprise and integrating anger and disgust improved the model performance. Li et al. modeled the label dependency by assuming a single sentence was likely to have similar labels such as hate and anger, and the context dependency of sentences by assuming that sentences in the same context were likely to have the same label using a factor graph (Li et al., 2015). The model using two neighboring sentences rather than the document or paragraph as the context showed the best performance.

Although it was a study on review documents, Zhang et al. formulated the sentiment analysis problem as a sentence sequence labeling problem (Zhang et al., 2014). They proposed a method to identify emotion labels (positive, negative, and neutral) by CRF (Lafferty et al., 2001). From the experimental results of active learning, the method labeling a document with the smallest average probability of the first half of the sentences in the document most effectively improved the model performance.

In recent years, in the field of natural language processing, approaches for fine-tuning a language model pre-learned with huge unlabeled text data in downstream tasks have been attracting attention (Qiu et al., 2020). BERT (Devlin et al., 2019) is a typical method of pre-training language representations. The effectiveness of the method using BERT has also been confirmed in the sentiment analysis task (Hoang et al., 2019; Sun et al., 2019; Xu et al., 2019). However, these models identify the emotion of an entire review or a single sentence. When considering the classification of emotions in each sentence of a news article, it is necessary to consider not only each sentence but also the contextual information around them.

Most conventional studies on sentiment analysis for news articles classify emotions at a finer granularity from the readers' perspective. In this study, we adopt three classes of "positive," "negative," and "neutral" for the granularity of the classification that can be agreed upon by many listeners. Similar to Zhang et al., we formulate the problem of identifying emotion labels for each sentence of news articles as a sentence sequence labeling problem.

3 News article corpus with emotion labels annotated for the sentences

We classified emotions in news content as "positive," "negative," or "neutral," and constructed a dataset for machine learning by annotating these emotion labels for each sentence in a news article. A positive label is annotated when the subject of the sentence considering the context is good or indicates that the subject is heading in a good direction. Examples include social contribution, market expansion, and acquisition of interests. A negative label is annotated when the subject of the sentence considering the context is bad or indicates that the subject is heading in a bad direction. Examples include a decline in business, acts of dishonesty, incidents, and accidents. A neutral label is annotated when neither positive nor negative content is included. Articles containing sentences with both positive and negative content were excluded in this study.

Annotation was performed by a web news clipping expert for news articles with 5 to 20 sentences in the Nihon Keizai Shimbun. The annotator was presented with lists of news articles ranked by category using a rule-based approach (see Section 5.1). The annotator annotated high ranked articles (news expected to be positive), middle ranked articles (news expected to be neutral), and low ranked articles (news expected to be negative) so that they were evenly distributed into the list. To cover various topics, we instructed the annotator to avoid annotating similar topics as much as possible. In the annotation work, first, the annotator read the title of the news article and checked the summary of the article. Next, the annotator assigned an emotion label to the title to understand the emotional tendency of the whole article. After reading all sentences of the article, the annotator assigned emotion labels to each sentence beginning from the first sentence.

Table 2 shows the total number of each emotion label annotated for titles and sentences by news category. The number of positive and negative annotated labels were almost equal, but there were fewer neutral labels.

Table 2: Number of emotion labels in titles and sentences for each news category

	#title			#sentence		
	Positive	Neutral	Negative	Positive	Neutral	Negative
Sports	65	10	71	555	399	564
Technology	91	8	93	776	519	908
Business	83	3	87	685	381	942
Markets	42	5	58	326	165	514
Economy	55	6	62	492	383	455
International	48	3	57	411	216	457
Society	114	4	98	860	559	699
Local	68	1	66	744	207	553

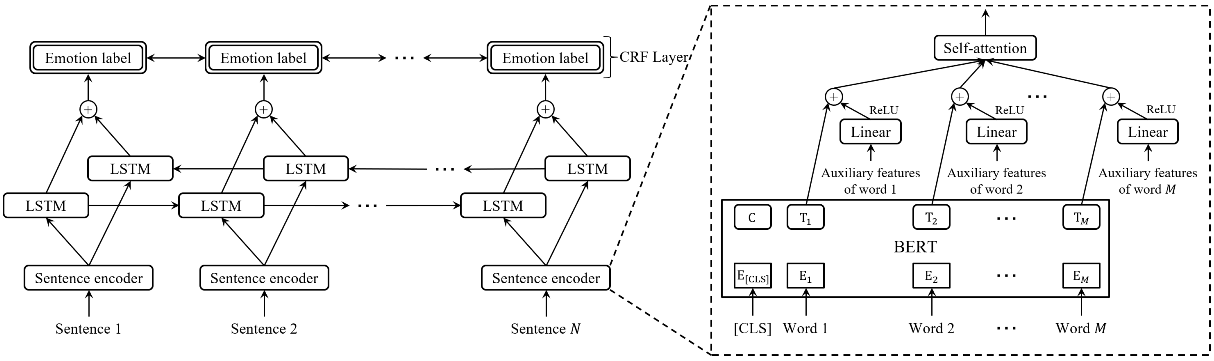

Figure 1: **BERT+_SA_BiLSTM-CRF** : Self-attention is calculated for the representation that combines the embedding of the top layer of BERT corresponding to each word and the embedding of the auxiliary features of each word. Obtained sentence vectors are given to BiLSTM-CRF, and the emotion labels of each sentence are estimated.

4 Proposed model

We formulated the emotion label identification problem as a sequence labeling problem. Figure 1 shows a schematic diagram of the proposed model. BERT (Devlin et al., 2019), which is a model based on the Transformer (Vaswani et al., 2017) encoder, is used as the word encoder. Self-attention (Lin et al., 2017) is calculated for the representation that combines the embedding of the top layer of BERT corresponding to each word and the embedding of the auxiliary features of each word. The obtained sentence vectors are given to BiLSTM-CRF (Lample et al., 2016), and the emotion labels of each sentence are estimated. At the time of decoding, the labels are estimated by the Viterbi algorithm.

The following information is used as auxiliary features: morphological information (part of speech, inflectional form, inflected type, category, domain) of JUMAN++ (Ver.1.02)[1] (Morita et al., 2015; Tolmachev et al., 2018), named entity classes and types of dependencies obtained by applying KNP (4.19)[2] (Kawahara and Kurohashi, 2006), distance from the top node of the dependency tree to the clause that contains the target word, the number of clauses whose destination is the clause that contains the target word, TF, IDF, TF-IDF, whether the word is included in the range of the corner bracket, clause position from the beginning of the sentence, position of the sentence in the article, position of the paragraph in the article, news category of the article, emotion polarity value of the word (using the “Japanese Sentiment Polarity Dictionary”[3] (Kobayashi et al., 2004; Higashiyama et al., 2008), the “Semantic Orientations of Words”[4] (Takamura et al., 2005), the polarity dictionary included in “Models for Opinion Extraction Tool”[5]), and whether the word is a polarity inversion expression (using the reverse expression dictionary included in “Models for Opinion Extraction Tool”).

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN++

[2] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

[3] https://www.cl.ecei.tohoku.ac.jp/index.php?Open%20Resource%2FJapanese%20Sentiment%20Polarity%20Dictionary

[4] http://www.lr.pi.titech.ac.jp/˜takamura/pndic_en.html

[5] https://alaginrc.nict.go.jp/li-outline.html#C-3

5 Experiments

5.1 Experimental setup

We evaluate the proposed model using the constructed dataset. We used the pre-trained BERT model published by Kyoto University[6]. This BERT model trained $\text{BERT}_{\text{BASE}}$ (Devlin et al., 2019) by inputting text applied to morphological analysis using JUMAN++[7] (Morita et al., 2015; Tolmachev et al., 2018) and BPE (Byte Pair Encoding)[8] (Sennrich et al., 2016) for all Japanese Wikipedia articles. The dimensions of the hidden layer of BiLSTM and linear layer were set to 128, and Adam was used for the optimizer. The macro F_1-measure (Chinchor, 1992) and overall accuracy were adopted as evaluation metrics. The evaluation was performed by the ten-fold cross validation where the dataset was divided into training set (90%) and test set (10%) for each news category. We compared the proposed model with following two types of methods.

Baselines 1 : Sentence classification methods

Random : A model that randomly selects a label.

Mode : A model that selects the most frequent labels (i.e., negative) in the dataset.

Rule-best : A model where the positive, neutral, and negative thresholds are adjusted to achieve the highest performance in the rule-based method. In the rule-based method, the emotion polarity value of a sentence is calculated according to the occurrence frequency of positive words and negative words by considering the polarity inversion using a word emotion polarity dictionary.

SVM : An SVM model of a linear kernel trained using the bag-of-words of sentence words as features.

BERT : A model that adds a linear layer on the top layer of BERT corresponding to [CLS] and applies Softmax.

BERT_SA+ : A model that applies Softmax to the vector obtained by calculating Self-attention for the combination of the embedding of the top layer of BERT corresponding to each word and the embedding of the auxiliary features of each word.

Baselines 2 : Sequence labeling methods

BiLSTM-CRF : A model that inputs the bag-of-words of sentence words into BiLSTM-CRF.

BERT_BiLSTM-CRF : A model that inputs the embedded representations of the top layer of BERT corresponding to [CLS] into BiLSTM-CRF.

BiLSTM+_SA_BiLSTM-CRF : A model that inputs vectors obtained by calculating Self-attention for a combination of the output vector of the hidden layer of the BiLSTM that inputs sentence words and the embedding of the auxiliary features of each word into BiLSTM-CRF.

5.2 Experimental results

The proposed model had the best performance (Table 3). Models using the embedded representations of BERT outperformed the model using the bag-of-words as word features. Additionally, the sequence labeling models outperformed the sentence classification models. Furthermore, the model performance could be improved by considering the auxiliary features.

Table 4 shows the values of each evaluation metric calculated by news category for BERT+_SA_BiLSTM-CRF. The “local” category had the best results. This is attributed to the prevalence of news with easy-to-understand tones in the “local” category such as incidents, accidents, and efforts to revitalize the region. In addition, neutral labels had a lower estimation performance than other labels regardless of news category. This is because even if a sentence contains positive or negative expressions, it may be neutral depending on the context.

[6] http://nlp.ist.i.kyoto-u.ac.jp/nl-resource/JapaneseBertPretrainedModel/Japanese_L-12_H-768_A-12_E-30_BPE.zip

[7] https://github.com/ku-nlp/jumanpp

[8] https://github.com/rsennrich/subword-nmt

Table 3: Performance of the models

Task	Model	Macro F_1-measure	Overall accuracy
Sentence classification	Random	0.340	0.345
	Mode	0.183	0.380
	Rule-best	0.578	0.608
	SVM	0.608	0.645
	BERT	0.685	0.722
	BERT+_SA	0.700	0.734
Sequence labeling	BiLSTM-CRF	0.670	0.723
	BERT_BiLSTM-CRF	0.738	0.774
	BiLSTM+_SA_BiLSTM-CRF	0.695	0.737
	BERT+_SA_BiLSTM-CRF	**0.773**	**0.805**

Table 4: Performance of BERT+_SA_BiLSTM-CRF for each news category

	F_1-measure				Overall accuracy
	Positive	Neutral	Negative	Macro average	
Sports	0.821	0.572	0.819	0.737	0.759
Technology	0.822	0.603	0.876	0.767	0.793
Business	0.856	0.531	0.898	0.762	0.820
Markets	0.831	0.623	0.878	0.777	0.825
Economy	0.828	0.657	0.818	0.768	0.777
International	0.807	0.536	0.828	0.724	0.759
Society	0.838	0.633	0.882	0.784	0.801
Local	0.933	0.638	0.942	0.837	0.898

6 Active learning

In the human-in-the-loop machine learning framework (Munro, 2019), the model performance can be efficiently improved by preferentially annotating articles with low confidence.

6.1 Confidence and accuracy

The confidence of labeling sentences in an article was defined as the value obtained by dividing the score from Viterbi decoding by the number of sentences. The dataset was divided into a training set (75%) and a test set (25%), and BERT+_SA_BiLSTM-CRF was trained. Figure 2 shows a scatter plot of the accuracy calculated for each article in the test set and the normalized confidence so that the maximum value is 1. Articles with a higher confidence tended to have a higher accuracy. Pearson's product moment correlation coefficient was 0.576, indicating an appropriate correlation.

6.2 Human-in-the-loop machine learning

The model performance and efficiency should be improved by preferentially annotating articles with low confidence. We evaluated the change in accuracy by repeating the following three steps:

(1) Apply the trained model to news articles with unknown labels that are not included in the training set and rank them in ascending confidence for each news category.

(2) Select the articles with the least confidence that match the annotation condition one-by-one for each news category and annotate the emotion labels.

(3) Add the annotated articles to the training set, retrain the model, and evaluate the performance using the test set.

Figure 3 shows an image of this annotation cycle. As a comparison, we also employed a model trained by adding data annotated in randomly selected articles for each news category regardless of the confidence.

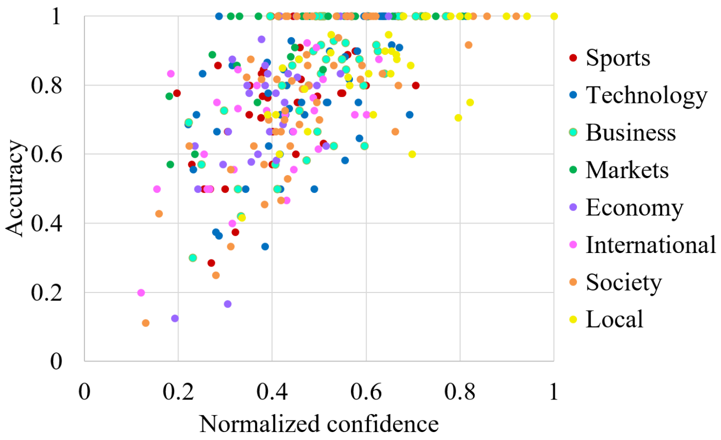

Figure 2: Scatter plot of the normalized confidence and accuracy in the test set.

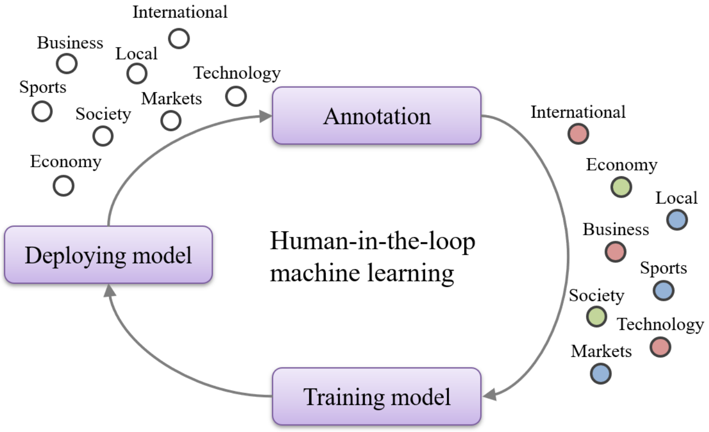

Figure 3: Human-in-the-loop machine learning. (1) Deploying model: Apply the trained model to news articles with unknown labels that are not included in the training set and rank them in ascending confidence for each news category. (2) Annotation: Select the articles with the least confidence that match the annotation condition one-by-one for each news category and annotate the emotion labels. (3) Training model: Add the annotated articles to the training set, retrain the model, and evaluate the performance using the test set.

6.3 Experimental results

Figure 4 plots the change in accuracy according to the number of active learning loops using the least confidence method and the random sampling method. The model using data annotating articles with a low confidence improved performance more efficiently than that using data annotated from randomly selected articles.

Figure 5 shows the error tendency for the before-the-loop model that had an accuracy of 0.794. Figure 6 shows the error tendency for the human-in-the-loop model after the five loops that had an accuracy of 0.809. Human-in-the-loop reduced the estimation errors of positive sentences as neutral, but it increased the errors of neutral sentences as positive, indicating that it is more difficult to distinguish between positive and neutral. In emotional speech synthesis, it is more critical when positive sentences are mistakenly identified as negative and vice versa. Such critical errors in both models were not more than 15%.

Table 5 shows an example of a news article where all the emotion labels were correctly estimated in the human-in-the-loop model. Table 6 shows an example of a news article where the emotion labels of some sentences were incorrectly estimated in the human-in-the-loop model. The model judged sentences with the content “the registered trademark is canceled” as negative. However, careful reading of the sentences

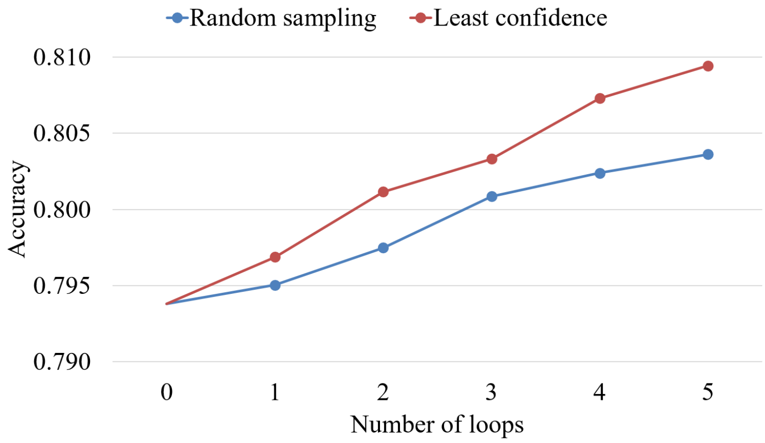

Figure 4: Number of loops of active learning and accuracy

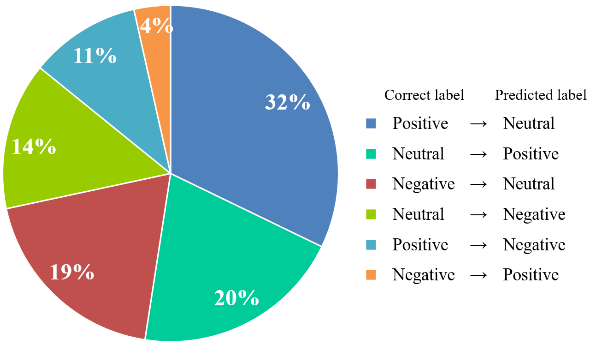

Figure 5: Error tendency in the before-the-loop model

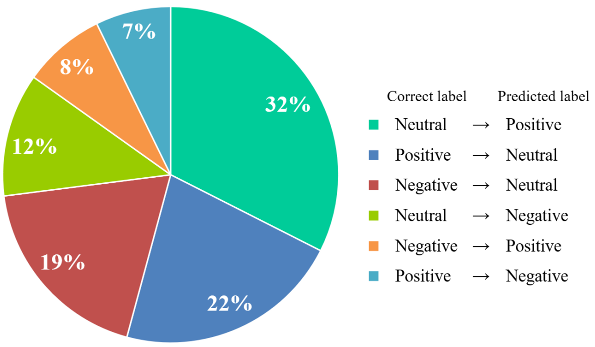

Figure 6: Error tendency in the human-in-the-loop model after the five loops

revealed that "a trademark registered without permission is canceled," can be considered positive content. One method to correct such errors is to create a model that learns the interpretation according to the context by increasing the amount of learning data. Another method is to introduce a mechanism into the model that can learn polarity inversion such that if the content is negative in a particular context, it becomes positive when the context changes.

7 Conclusion

To enable emotional speech synthesis based on news content in a spoken dialogue system, we constructed a dataset for machine learning by annotating the emotion labels for each sentence in a news article. In addition, we proposed a model that identifies the emotion label of each sentence, and evaluated its effectiveness using the constructed dataset. The model performance can be efficiently improved by preferentially annotating articles with low confidence in the human-in-the-loop machine learning framework.

In the future, we will develop a speech synthesis system that can control the emotional parameters using the emotion label estimated by the proposed model. We will also confirm whether speaking with emotion promotes users' understanding in news delivery tasks.

Acknowledgements

This work was supported by Japan Science and Technology Agency (JST) Program for Creating STart-ups from Advanced Research and Technology (START), Grant Number JPMJST1912 "Commercialization of Socially-Intelligent Conversational AI Media Service".

Table 5: Examples of a news article where the emotion labels of all sentences were correctly estimated

Sentence	Correct label	Predicted label
米航空機大手ボーイングは8日、2018年の商業用航空機の出荷が過去最高の806機だったと発表した。 US aircraft giant Boeing announced on the 8th that its shipment of commercial aircraft in 2018 was the highest ever, 806.	Positive	Positive
米国やアジアで小型機「737MAX」の需要が伸び、過去最高だった17年を6%上回った。 The demand for the small machine “737 MAX” grew in the US and Asia, surpassing the record high of 2017 by 6%.	Positive	Positive
部品調達の遅れなどで810〜815機としていた計画には届かなかった。 The plan for 810 to 815 new aircraft could not be reached due to delays in the supply chain.	Negative	Negative
格安航空会社を中心に単通路の737シリーズの需要が伸び、同シリーズの出荷が580機と全体の7割を占めた。 The demand for the single-passage 737 series grew, centering on low-cost airlines, which received 580 aircraft, accounting for 70% of the total.	Positive	Positive
18年の受注は893機だった。 There were 893 orders for aircrafts in 2018.	Neutral	Neutral
昨年10月にインドネシアで737MAXの墜落事故が発生したが、受注に目立った影響は出ていない。 In October of last year, a 737 MAX crashed in Indonesia, but there was no noticeable effect on orders.	Neutral	Neutral

Table 6: Example of a news article where the emotion labels of some sentences were incorrectly estimated

Sentence	Correct label	Predicted label
人気が高い鹿児島県の芋焼酎が中国での商標を無断登録されていた問題で、「森伊蔵」「伊佐美」の2銘柄について中国商標局が登録の取り消しを決定したことが2日分かった。 It was revealed on the 2nd that the Chinese trademark office revoked the registration of two brands, “Mori Izo” and “Isami” due to the problem that potato shochus in Kagoshima prefecture, which are very popular, were registered without a trademark in China.	Positive	Negative
2つの商標は関係のない福岡県の会社が商標登録をしていたが、それぞれの焼酎を製造する森伊蔵酒造と甲斐商店が取り消しを求めていた。 A company in Fukuoka prefecture, which is unrelated to the two trademarks, registered the trademark. However, Mori Izo Shuzo and Kai Shoten, which manufacture their respective shochus, were seeking cancellation.	Neutral	Negative
今回の決定を受け、両社とも中国商標局に対して商標登録を申請した。 Following this decision, both companies applied for trademark registration with the Chinese Trademark Office.	Neutral	Neutral
両社はこれまでも無断出願に対して中国当局に異議申し立てをしてきたが、先に申請した者に権利を与える先願主義の壁に阻まれて認められていなかった。 The two companies challenged the Chinese authorities for unapproved applications. However, they have not been admitted, hampered by the first-to-file barrier that gives rights to earlier applicants.	Negative	Negative
今回は一定期間使用がない場合に商標を取り消す制度を活用、「3年間不使用」だったとして請求が認められたという。 This time, they used the system to cancel the trademark because there is a rule about non-use for a certain period, and the request was granted as “not used for 3 years.”	Positive	Negative

References

Jerome R. Bellegarda. 2011. A data-driven affective analysis framework toward naturally expressive speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1113–1122.

Maren Bennewitz, Felix Faber, Dominik Joho, and Sven Behnke. 2007. Fritz - a humanoid communication robot. In *Proceedings of the 16th IEEE International Conference on Robot & Human Interactive Communication*, pages 1072–1077.

Plaban Kr. Bhowmick, Anupam Basu, Pabitra Mitra, and Abhishek Prasad. 2009. Multi-label text classification approach for sentence level news emotion analysis. In *Proceedings of the 3rd International Conference on Pattern Recognition and Machine Intelligence*, pages 261–266.

Plaban Kumar Bhowmick, Basu Anupam, and Mitra Pabitra. 2010. Classifying emotion in news sentences: When machine classification meets human classification. *International Journal on Computer Science and Engineering*, 2(1):98–108.

Marcela Charfuelan and Ingmar Steiner. 2013. Expressive speech synthesis in mary tts using audiobook data and emotionml. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 1564–1568.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1):321–357.

Yuya Chiba, Takashi Nose, Taketo Kase, Mai Yamanaka, and Akinori Ito. 2018. An analysis of the effect of emotional speech synthesis on non-task-oriented dialogue system. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 371–375.

Nancy Chinchor. 1992. Muc-4 evaluation metrics. In *Proceedings of the 4th conference on Message understanding*, pages 22–29.

Ciptadi and Abba Suganda Girsang. 2019. Emotion classification based on public opinion analysis on online news. *International Journal of Scientific & Technology Research*, 8(6):176–182.

Francesco De Comité, Rémi Gilleron, and Marc Tommasi. 2003. Learning multi-label alternating decision trees from texts and data. In *Proceedings of the 3rd International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 35–49.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

Dipankar Das and Sivaji Bandyopadhyay. 2009. Analyzing emotion in blog and news at word and sentence level. In *Proceedings of the 4th Indian International Conference on Artificial Intelligence*, pages 1402–1414.

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*, pages 4171–4186.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442.

D. Govind and S. R. Mahadeva Prasanna. 2013. Expressive speech synthesis: A review. *International Journal of Speech Technology*, 16(2):237–260.

Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. 2008. Learning sentiment of nouns from selectional preferences of verbs and adjectives. In *Proceedings of the 14th Annual Meeting of the Association for Natural Language Processing*, pages 584–587. (in Japanese).

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 187–196.

Ronglei Hu, Lu Rui, Ping Zeng, Lei Chen, and Xiaohong Fan. 2018. Text sentiment analysis: A review. In *Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications*, pages 2283–2288.

Katsuki Inoue, Sunao Hara, Masanobu Abe, Nobukatsu Hojo, and Yusuke Ijima. 2017. An investigation to transplant emotional expressions in dnn-based tts synthesis. In *Proceedings of the 9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1253–1258.

Jumayel Islam, Robert E. Mercer, and Lu Xiao. 2019. Multi-channel convolutional neural network for twitter emotion and sentiment recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1355–1365.

Kazuhiko Iwata and Tetsunori Kobayashi. 2011. Conversational speech synthesis system with communication situation dependent hmms. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 113–123.

Igor Jauk, Jaime Lorenzo-Trueba, Junichi Yamagishi, and Antonio Bonafonte. 2018. Expressive speech synthesis using sentiment embeddings. In *Proceedings of the 19th Annual Conference of the International Speech Communication*, pages 3062–3066.

Jasmine Kaur and Parminder Singh. 2015. Review on expressive speech synthesis. *International Journal of Electrical, Electronics and Computer Systems*, 3(10):24–28.

Daisuke Kawahara and Sadao Kurohashi. 2006. A fully-lexicalized probabilistic model for japanese syntactic and case structure analysis. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 176–183.

D. K. Kirange and R. R. Deshmukh. 2012. Emotion classification of news headlines using svm. *Asian Journal of Computer Science and Information Technology*, 5(2):104–106.

Nozomi Kobayashi, Kentaro Inui, Yuji Matsumoto, Kenji Tateishi, and Toshikazu Fukushima. 2004. Collecting evaluative expressions for opinion extraction. In *Proceedings of the 1st International Joint Conference on Natural Language Processing*, pages 596–605.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Shoushan Li, Lei Huang, Rong Wang, and Guodong Zhou. 2015. Sentence-level emotion classification with label and context dependence. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1045–1053.

Xin Li, Haoran Xie, Yanghui Rao, Yanjia Chen, Xuebo Liu, Huan Huang, and Fu Lee Wang. 2016. Weighted multi-label classification model for sentiment analysis of online news. In *Proceedings of the 2016 International Conference on Big Data and Smart Computing*, pages 215–222.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2007. What emotions do news articles trigger in their readers? In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 733–734.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 220–226.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *Proceedings of the 5th International Conference on Learning Representations*, pages 1–15.

June Ling, Ong Hui, Gan Keng Hoon, Wan Mohd Nazmee, and Wan Zainon. 2017. Effects of word class and text position in sentiment-based news classification. In *Proceedings of the 4th Information Systems International Conference*, pages 77–85.

Huanhuan Liu, Shoushan Li, Guodong Zhou, Chu-Ren Huang, and Peifeng Li. 2013. Joint modeling of news reader's and comment writer's emotions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 511–515.

Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. 2015. Morphological analysis for unsegmented languagesusing recurrent neural network language model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297.

Robert Munro. 2019. *Human-in-the-loop machine learning*. Manning.

Takashi Nose and Takao Kobayashi. 2013. An intuitive style control technique in hmm-based expressive speech synthesis using subjective style intensity and multiple-regression global variance model. *Speech Communication*, 55(2):347–357.

Sandip S. Patil and Asha P. Chaudhari. 2012. Classification of emotions from text using svm based opinion mining. *International Journal of Computer Engineering & Technology*, 3(1):330–338.

John F. Pitrelli, Raimo Bakis, Ellen M. Eide, Raul Fernandez, Wael Hamza, and Michael Picheny. 2006. The ibm expressive text-to-speech synthesis system for american english. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1099–1108.

Long Qin, Zhen-Hua Ling, Yi-Jian Wu, Bu-Fan Zhang, and Ren-Hua Wang. 2006. Hmm-based emotional speech synthesis using average emotion model. In *Proceedings of the 5th International Symposium on Chinese Spoken Language Processing*, pages 233–240.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. In *arXiv:2003.08271*, pages 1–28.

Philipp Robbel, Mohammed E. Hoque, and Cynthia Breazeal. 2009. An integrated approach to emotional speech and gesture synthesis in humanoid robots. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, pages 1–4.

Marc Schröder. 2001. Emotional speech synthesis: A review. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, pages 561–564.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725.

Mostafa Al Masum Shaikh, Antonio Rui Ferreira Rebordao, Keikichi Hirose, and Mitsuru Ishizuka. 2009. Emotional speech synthesis by sensing affective information from text. In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6.

R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4700–4709.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pages 70–74.

B. Sudhakar and R. Bensraj. 2014. An efficient sentence-based sentiment analysis for expressive text-to-speech using fuzzy neural network. *Engineering and Technology*, 8(3):378–386.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–385.

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 133–140.

Hiroaki Takatsu, Ishin Fukuoka, Shinya Fujie, Yoshihiko Hayashi, and Tetsunori Kobayashi. 2018. A spoken dialogue system for enabling information behavior of various intention levels. *Journal of the Japanese Society for Artificial Intelligence*, 33(1):1–24. (in Japanese).

Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. 2018. Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations)*, pages 54–59.

Alexandre Trilla and Francesc Alías. 2013. Sentence-based sentiment analysis for expressive text-to-speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(2):223–233.

Eva Vanmassenhove, João P. Cabral, and Fasih Haider. 2016. Prediction of emotions from text using sentiment analysis for expressive speech synthesis. In *Proceedings of 9th ISCA Speech Synthesis Workshop*, pages 21–26.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Łukasz Kaiser. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing System*, pages 6000–6010.

Jenq-Haur Wang and Hao-Ying Liu. 2017. Discovering reader's emotions triggered from news articles. In *Proceedings of the 4th Multidisciplinary International Social Networks Conference*, pages 1–7.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2324–2335.

Hongwu Yang, Weizhao Zhang, and Pengpeng Zhi. 2018. A dnn-based emotional speech synthesis by speaker adaptation. In *Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 633–637.

Da Yin, Xiao Liu, Xiuyu Wu, and Baobao Chang. 2019. A soft label strategy for target-level sentiment classification. In *Proceedings of the 10th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 6–15.

Yuan Zhang and Yue Zhang. 2019. Tree communication models for sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3518–3527.

Kunpeng Zhang, Yusheng Xie, Yi Yang, Aaron Sun, Hengchang Liu, and Alok Choudhary. 2014. Incorporating conditional random fields and active learning to improve sentiment identification. *Neural Networks*, 58:60–67.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis : A survey. In *arXiv:1801.07883*, pages 1–34.