An Investigation of Sentiment Recognition with Error Prone Multimodal Language Sequences

Wei Yang and Jun Ogata
Artificial Intelligence Research Center, National Institute
of Advanced Industrial Science and Technology (AIST)
wei.yang@aist.go.jp; jun.ogata@aist.go.jp

Abstract

In this work, we propose to investigate the sentiment recognition accuracy using unaligned multimodal language sequences including error prone textual modality. For that, we firstly construct an ASR system using a publicly available corpus based on Transformer architecture that relying on self-attention mechanism. Then, for automatically creating the textual data for an existing audio dataset, we perform speech-to-text on the audio part of a multimodal dataset. Finally, we perform sentiment recognition experiments with multi-modality language sequences including the textual features that extracted based on automatically created error prone textual data instead of utilizing manually created transcripts. We estimate that the word error rate (WER) of our ASR system is 4.87% in the case of SentencePiece based subword segmentation applied for the textual data in training process. We obtain the best accuracy of sentiment recognition experiments conducted on error prone multimodal language sequences with 70%.

1 Introduction

Emotion and sentiment recognition is crucial as an application of artificial intelligence (AI) in human-computer interaction (HCI), and it has become an overlapping research between psychology and computer science (CS). But as we know, it is difficult to understand human language by AI, especially through only one behavior (unimodality) as human language is complicated due to human's dynamic multi-modality nature during face-to-face communication. In other words, it is always dynamic and variable language sequences including verbal behavior (natural language/textual modality) and nonverbal behaviors (acoustic modality and visual modality). On the basis of experience, more exact intentions should be more easier to capture by considering natural language and nonverbal cues at the same time (multimodality/multimodal).

The crucial process of emotion and sentiment recognition should be helpful by the use of multimodal data driven recognition engine using state-of-the-art AI techniques. Thus, for both training process and evaluating how reliable a multimodal engine is, creating large-scale labelled multimodal datasets in different languages on all domains is essential and extremely important. Some research institutions have consumed a lot of time and effort to construct publicly available dataset, such as CMU-MOSI [20], CMU-MOSEI [1] and IEMOCAP [3]. We should give many thanks to these works. But we also have to notice that manually constructing multimodal dataset in different modalities is time consuming and cost. It is also difficult to apply one system on all scenarios in real life. Consequently, as an application of emotion or sentiment recognition for addressing real world problems with real time-series data becomes a challenge in research and development in artificial intelligence field.

From the previous work, we can see that much effort also has been made to modelling sentiment or emotion recognition from multiple modalities [15, 17]. In this work, we focus on the "Multimodal Transformer (MulT)" model [15] that used for emotion and sentiment analysis using multimodal datasets. The main advantage of "MulT" is that it can address "unaligned" problem for the time-series (sequences) data in different modalities, as well as the problem of long-range dependencies across modalities by using crossmodal attention mechanism (*"repeated reinforce one modality's features with the features from the other modalities"* [15]) to adapt streams from one modality to another. Actually, "MulT" is an extended end-to-end model based on standard Transformer architecture [16]. "MulT" can allow us to obtain state-of-the-art results in emotion and sentiment recognition field for several benchmarks. We refer the read to [15] and [16] for a more detail explanation of the overall architecture of "MulT" and the standard Transformer network. In our work, we only use CMU-MOSI as the multimodal dataset for sentiment recognition experiments.

In this paper, we construct an ASR system based on a standard Transformer architecture with a freely available dataset. We then perform speech-to-text experiments

relying on the pre-trained ASR systems for a publicly available multimodal language dataset. Finally, we extract textual features in word embeddings from those automated transcriptions combine with acoustic and visual features to perform sentiment recognition experiments. We obtain a best recognition accuracy with 70% by utilizing this kind of error prone multimodal language sequences in three modalities. We also compare the result with the accuracy of using only two modalities (acoustic and visual modalities, without textual modality). Three modalities, in spite of containing error prone textual features, still allow us to obtain much better results.

2 Sentiment Recognition with Error Prone Multimodal Dataset

2.1 Construction of ASR System

Some researchers have tried to work on speech representation learning approach for speech emotion recognition (SER) task through the use of automatic speech recognition (ASR) system [2, 19]. Different from these works, in our work, we directly use ASR system to automatically convert audio to text (transcripts) so as to obtain a multimodal language dataset that containing three modalities: language/textual modality, acoustic modality and visual modality. In speech recognition task, recurrent sequence-to-sequence (seq2seq) model using encoder-decoder network have achieved significant word error rate (WER) by introducing different techniques [21]. There also exist several open tookits that used for speech recognition task, for instance, Kaldi [13] and ESPnet [18]. In our experiments, we construct a Pytorch implementation of ASR system refer to ESPnet using a standard Transformer architecture. We do not apply speed perturbation [8] and CMVN (Cepsturm Mean and Variance Normalization) for experimental data preparation and feature extraction, but apply "SpecAugment" [11] for data argument.

In our experiments, we use Transformer architecture to train our ASR models. Thus, for position representations, we use absolute position encodings in self-attention mechanism that introduced in [16]. As using subword information has already become crucial to the improvement of the tasks in natural language processing (NLP), such as machine translation (MT) or the task we are doing here, speech recognition. Some previous work shows that using "subword" is an effective way for addressing unseen word or rare word issues, alleviating the open vocabulary problems [14], so as to obtain a stronger neural machine translation (NMT) system. Thus, we perform subword segmentation to further segment words into "subword" units for preparing the textual data used for training ASR systems.

In our experiments, we apply two different subword implementations on the textual data. One is "BPE" subword segmentation which means the segmenter relied on byte-pair-encoding (BPE) compression algorithm [5]. Another one is "SentencePiece (sp)" subword segmentation. "SentencePiece (sp)" supports two segmentation algorithms: "BPE" and "unigram language model" [9], here, for the case of "SentencePiece (sp)" we use "unigram language model". Table 1 shows the examples of word/subword segmentation results for the same text from our experimental dataset.

Word/sp/BPE based segmentation	Textual data for ASR
Word	one could have eaten a meal off the ground without overbrimming the proverbial peck of ...
SentencePiece (sp)	_one _could _have _eaten _a _meal _off _the _ground _without _over b ri m m ing _the _proverb ial _pe ck _of ...
BPE	one could have eaten a meal off the ground without over@@ brimming the proverbial peck of ...

Table 1: Examples of word/subword based segmentation for the same text from our experimental dataset.

Textual data (transcripts) are automatically created using our Transformer based ASR systems as described above based on "BPE" and "SentencePiece (sp)" respectively. We evaluate and compare the WER of the ASR systems for both test set including in ASR construction experiments and the audio data with their manual transcripts from an existing multimodal dataset (CMU-MOSI). We then choose to use the transcripts of the audio dataset that created from the better ASR system with lower WER as the third modality information act on multimodal sentiment recognition. The textual features (word embeddings) will be extracted afterwards for preparing the final experimental data used in sentiment recognition experiments.

2.2 Using Error Prone Multimodal Language Sequences in Sentiment Recognition

We propose three protocols to perform sentiment recognition experiments with multi-modality language sequences using "MulT". We compare these three protocols with each other, also with the results given in [15]. The first protocol (aligned) is that we train the sentiment recognition model based on our error prone multimodal language sequences are aligned to the same length. The second protocol (unaligned) is as follows. The length of the textual sequences is the same as the length obtained in the first protocol, but we keep the original length of the audio and visual features without any word-segmented alignment. The third protocol (unaligned but language only) is that we train the sentiment recognition model

using only the textual features that reinforced by audio and visual features.

3 Experiments

3.1 ASR and Sentiment Recognition Experimental Datasets

The English dataset used in our ASR system construction is a publicly available ASR corpus: LibriSpeech [10]. LibriSpeech is a publicly available ASR corpus[1] for speech recognition based on public domain audio books in English. We train Transformer based ASR models on the LibriSpeech corpus with 960 hours (train-960) of speech samples at 16kHz, whose contains three subsets (train-clean-100, train-clean-360 and train-other-500). We evaluate our ASR systems with "test-clean" dataset.

For our sentiment recognition experiments, we use the processed CMU-MOSI experimental dataset[2] which consisting multimodal features (language (L), audio (A) and vision (V)) for 2,199 short monologue video clips. The multimodal features are extracted from the textual, acoustic and visual multimodal dataset using Glove word embeddings (glove.840B.300d) [12], COVAREP [4] and Facet [6] respectively.

3.2 ASR System Construction and Speech-to-text Tasks

We train our ASR systems for 72 hours on 4GPUs. We average the last 8, 10 and 20 saved models for decoding the test set and report the best one. The input acoustic features are 40-dimensional filter banks. In the training stage, we used Adam optimizer [7] with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\varepsilon = 10^{-9}$ and varied the learning rate over the course of training. We used warmup_step=12000, we set the residual dropout with 0.1. Number of blocks for encoder is 12 and number of blocks for decoder is 6. The model dimension d_model=320 and the number of head is set with 4. The feed-forwad inner dimension is 1280. We use "GLU (Gated Linear Unit)" as the active function type in our experiments. The textual data of "train-960" dataset is used to learn a language model using Transformer as well. The pre-trained language model is also used in final prediction.

As the evaluation results shown in Table 2. Subword size used in our experiments for "sp" and "BPE" are 5,000 and 32,000 respectively. Vocabulary size are 5,000 and 31,398 obtained from subword segmentation results correspondingly. Our pre-trained Transformer based ASR systems allow us to obtain 4.87% and 5.25% in WER on LibriSpeech's "test-clean" dataset based on "SentencePiece (sp)" and "BPE" subword segmentation respectively. Meanwhile, we obtain 52.23% and 64.59% on the CMU-MOSI audio dataset, the reference is the original manual transcripts (textual data) in CMU-MOSI dataset. "SentencePiece (sp)" based subword segmentation obtain 12.36% improvement in WER over "BPE" subword segmentation on CMU-MOSI dataset while 0.38% relative improvement on the "test-clean" of LibriSpeech. The high scores in WER for CMU-MOSI dataset probably due to differences in domains and speaking styles.

Table 3 shows the samples of speech-to-test experiment results for CMU-MOSI dataset. We can see that some audio samples (e.g., id: OtBXNcAL_lE_18) cannot be recognized successfully using our ASR system (empty), maybe they are too short or so rapid to difficult to recognized. Some of the results do not match the reference because there contains "smacking lips" like stress information in the original manual transcripts (e.g., id: tmZoasNr4rU_10). There also exist some cases that ASR based transcripts are much more longer than their original manual transcripts in the dataset, and the ASR based transcripts are more reasonable according to their corresponding audio samples (e.g., id: 2WGyTLYerpo_44).

3.3 Construction of Multimodal Sentiment Recognition Systems

We perform the experiments with "aligned" and "unaligned" protocols almost the same as described in [15]. The hyperparameters of "MulT" use here are also the same as used in [15]. The only difference is that instead of using the extracted textual features given in the original CMU-MOSI processed experimental dataset, we replace them with our textual features. They are also extracted using the same "glove.840B.300d", but based on our automatically created error prone textual data (transcripts). Table 4 shows the evaluation results for the three protocols given in Sec. 2.2. Because of the error prone textual data, as expected, we did not obtain the same results in all evaluation metrics in comparison with the results obtained in [15]. But we compare the results with each other, we found that our results are on the contrary to the results obtained based on the original CMU-MOSI multimodal language sequences with manual textual data. In our experiments, "unaligned" has better results in comparison with "aligned" protocol. Meanwhile, reinforcing language (L) features with audio (A) and visual (V) features allow us to obtain the best result. This is to say, Multimodal Transformer with crossmodal attention module are so efficient for addressing the problems of long term dependencies across modalities, especially for "unaligned", even "error prone" nature for multimodal language sequences. We also perform the

[1] http://www.openslr.org/12

[2] http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/cmu-mosi/seq_length_50/

Dataset (train-960)	Model	Dataset (test)	Subword segmentation	WER (%)
LibriSpeech	Transformer	LibriSpeech	SentencePiece	**4.87**
			BPE	5.25
LibriSpeech		CMU-MOSI	SentencePiece	**52.23**
			BPE	64.59

Table 2: Evaluation results of speech-to-text experiments based on our Transformer ASR systems for both "test-clean" LibriSpeech test set and CMU-MOSI data. The scores in WER for two different subword segmentation strategies is given on the last column. The number in boldface shows the better results.

Utterance-id	by ASR or Manual	Text
tmZoasNr4rU_10	ASR	`maybe only five jokes made me laugh`
	Manual	`smacking lips maybe only 5 jokes made me laugh`
OtBXNcAL_lE_18	ASR	`(empty)`
	Manual	`and blah blah blah`
2WGyTLYerpo_44	ASR	`and it's not a bad actress she's actually pretty good and she's in everything ... (62 words in total)`
	Manual	`shes beautiful (it should be about 167 words in total)`

Table 3: Examples of speech-to-test experiment results for CMU-MOSI dataset. Here the subword segmentation strategy used is "SentencePiece" (after "SentencePiece" model decoding).

Metric	Acc_7	Acc_2	F1	MAE	Corr
(Word aligned) CMU-MOSI Sentiment					
ours	26.2	67.4	67.3	1.260	0.470
MulT in [15]	40.0	83.0	82.8	0.871	0.698
(Unaligned) CMU-MOSI Sentiment					
ours	26.8	68.9	68.8	1.198	0.485
MulT in [15]	39.1	81.1	81.0	0.889	0.686
(Unaligned, language only) CMU-MOSI Sentiment					
Only [V, A $\rightarrow$ L] (ours)	**30.1**	**70.1**	**70.1**	**1.186**	**0.485**
(Unaligned, two modalities) CMU-MOSI Sentiment					
Only [A + V] (ours) (A $\rightarrow$ V + V $\rightarrow$ A)	19.2	57.8	57.5	1.359	0.208

Table 4: Results for the "MulT" model based multimodal sentiment experiments on CMU-MOSI with our error prone aligned and unaligned multimodal language sequences. As the evaluation metrics, expect MAE is the lower the better, others are the higher the better. Acc_7: 7-class accuracy, Acc_2: binary accuracy, Corr: correlation of model's prediction with human, MAE: mean absolute error of the score.

experiments only with acoustic and visual modalities using "MulT". From the results, we can see that textual features is extremely important in sentiment recognition, in spite of these textual data are error prone and not perfect.

4 Conclusion

In this paper, we investigated that as an application of emotion or sentiment recognition for solving real life problems, in other words, if we prefer to do some individual projects, how about the sentiment recognition accuracy that perform with error prone multimodal language sequences in multiple modalities. We assume that we only have acoustic and visual data modalities, for having a multimodal dataset including textual modality, we tried to create textual data using an pre-trained ASR system. The results show that three-modality language sequences combine with a robust multimodal emotion or sentiment recognition system can allow us to obtain a better classification result in comparison with using only two modalities, in spite of the textual features in the language aspect suffers from the low quality of automated transcriptions. From our results, we also can conclude that textual modality is definitely an essential and important information provides us to understand human's behaviors and intents more exactly.

5 Acknowledgements

Part of this work was supported by Council for Science, Technology and Innovation, "Cross-ministerial Strategic Innovation Promotion Program (SIP), Big-data and AI-enabled Cyberspace Technologies". (Funding agency: NEDO)

References

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meet-*

ing of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246, 2018.

[2] Yonatan Belinkov and James Glass. Analyzing hidden representations in end-to-end automatic speech recognition systems. 2017.

[3] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[4] G. Degottex, John Kane, Thomas Drugman, T. Raitio, and Stefan Scherer. COVAREP–a collaborative voice analysis repository for speech technologies. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2014)*, pages 960–964, 2014.

[5] Philip Gage. A new algorithm for data compression. *C Users J.*, 12(2):23–28, 1994.

[6] iMotions. Facial expression analysis. 2017.

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. URL `http://arxiv.org/abs/1412.6980`.

[8] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. Audio augmentation for speech recognition. In *INTERSPEECH*, 2015.

[9] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. 2018.

[10] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE international Conference on Acoustics, Speech and Signal Processing (ICASSP2015)*, pages 5206–5210, 2015.

[11] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Proc. Interspeech 2019*, pages 2613–2617, 2019.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[13] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 2011.

[14] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, 2016.

[15] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. 2019.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS'17 Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, CA, USA, 2017.

[17] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. 2018.

[18] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. Espnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech*, pages 2207–2211, 2018.

[19] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. Speech Representation Learning for Emotion Recognition Using End-to-End ASR with Factorized Adaptation. In *Proc. Interspeech 2020*, pages 536–540, 2020.

[20] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *IEEE Intelligent Systems 31.6 (2016): 82-88*, 2016.

[21] Yu Zhang, William Chan, and Navdeep Jaitly. Very deep convolutional networks for end-to-end speech recognition. 2016.