Sequential Attention-based Detection of Semantic Incongruities from EEG While Listening to Speech

Shunnosuke Motomura[1], Hiroki Tanaka[1], Satoshi Nakamura[1]

***Abstract*— We propose a method with attention-based recurrent neural networks (ARNN) for detecting the semantic incongruities in spoken sentences using single-trial electroencephalogram (EEG) signals. 19 participants listened to sentences, some of which included semantically anomalous words. We recorded their EEG signals while they listened. Although previous detection approaches used a word's explicit onset, we used the EEG signals of the whole regions of each sentence, which made it possible to classify the correctness of the sentences without the onset information of the anomalous words. ARNN achieved 63.5% classification accuracy with a statistical significance above the chance level and also above the performances which includes onset information (50.9%). Our results also demonstrated that the attention weights of the model showed that the predictions depended on the feature vectors that are temporally close to the onsets of the anomalous words. *Clinical relevance*— This technique also can be applied to measuring people's traits of amnestic mild cognitive impairment such as Alzheimer's disease in terms of semantic impairment. *Keywords*- Single-trial EEG, human sentence processing, N400, attention-based recurrent neural networks**

I. INTRODUCTION

Human beings recognize semantic incongruities or ambiguities in sequences, e.g. in essays, utterances by a language learner, or system-generated sentences. To evaluate these incongruities, subjective evaluations are usually used. However, they can be affected by biases caused by subjective factors because of the difficulty of defining clear criteria for the evaluations or the interpretations of the meanings of words; there is also no assurance that answers are correct [1]. In this paper, we propose the method to detect semantic incongruities in spoken sentences for automatic real-time evaluations with EEG signals, which include the spontaneous signals of the neurons of brains from which we can acquire the high time-resolution information specific to a certain stimulus [2]. It also can be applied to measuring people's traits of amnestic mild cognitive impairment such as Alzheimer's disease in terms of semantic impairment [3].

N400, which is a component of event-related potentials (ERPs), can be observed in signals for sentences, including semantic anomalies [4]. Therefore, its size is correlated with a word's expectancy to a preceding context, i.e. cloze probability [4]. For the observation of ERP components, we must average the signals of multiple trials'. This step basically requires that at least 50 trials be averaged to observe significant ERP components [2]. For real-time online evaluation, we must consider single-trial level detection, which is very challenging due to the low signal-to-noise ratio of EEGs. Few works have studied the classification problems of single-trial EEG signals [5] and achieved 61.3% accuracy. One study attempted to detect semantic anomalies in spoken sentences [6] using sentences, some of which included semantically anomalous words placed in the third-word positions of sentences and extracted the EEG signals that correspond to the specific words for classification. Models using multilayer perceptrons showed 59.5% accuracy. However, since we showed that semantic incongruities are related to a word's expectancy to a preceding context, they must be tested for the classification methods using EEG signals, which uses the whole parts of sentences. We hypothesized that the EEG signals when participants recognized other words in the sentences might provide classification information and also we cannot know the onset (and its timing may be ambiguous, specifically the speech stimulus) where in the sentence semantic anomalies occur in real situations.

Methods have recently been proposed with deep neural network models to classify single-trial EEG signals [7] and some studies reported that recurrent neural networks (RNNs) handle sequential features well for EEG classifications [8]. Sequential attention mechanisms [9], which decide the importance at specific time areas of the signals for predictions can be powerful tools for such problems as the classification of sequential EEG signals [9], [10]. We can also analyze the parts of the sentences the attention mechanism focused on and the weights in the EEG signals of each sentence.

To the best of our knowledge, no attempts have studied how well sequential attention models perform EEG classification related to semantic comprehension. We propose a method with attention-based RNN models using the features of the EEG signals of the whole parts of each sentence for detecting semantic anomalies in speeches. The following are the three contributions of this paper: (1) We detected semantic anomalies with the EEG signals of the whole parts of individual sentence; (2) we evaluated the performances of attention-based RNNs for language-related EEG signals; (3) we analyzed the parts of the sentences the attention mechanism focused on and the weights in the EEG signals of each sentence.

II. METHOD

In this section, we describe our classification model and the experimental data collection for the single-trial detection

*This work was not supported by JSPS KAKENHI Grant Numbers JP17H06101, JP18K11437

[1]Shunnosuke Motomura, Hiroki Tanaka and Satoshi Nakamura are with Nara Institute of Science and Technology, Japan {motomura.shunnosuke.mj1, hiroki-tan, s-nakamura}@is.naist.jp

of the semantic incongruities in EEG signals.

A. Detection Model

We used bidirectional gated recurrent units (GRUs) as an RNN classifier and introduced a sequential attention mechanism for predicting incongruities from sequential inputs.

1) Gated Recurrent Units: A gated recurrent unit (GRU) [11], which is a kind of RNN with reset gates, update gates and hidden states at each time step, is one version of long short-term memory. In this paper, we used bidirectional GRUs for our classifications.

2) Attention-based Recurrent Neural Networks (ARNN): Since the signals at all the time points in the sentences are not equally useful for classifications, we used RNNs with an attention mechanism that can assign the importance scores at each time point and construct feature vectors with representations of the whole time regions [12]. We can calculate the attention weights at each time point as follows:

$$\alpha_t = \frac{exp(h_t^{\mathsf{T}} w)}{\sum_t exp(h_t^{\mathsf{T}} w)} \quad (1)$$

$$v = \sum_t \alpha_t h_t \quad (2)$$

where h_t is the RNN output at time t and w is a trainable attention vector. Thus, $h_t^{\mathsf{T}} w$ represents the importance at time t by measuring the similarity between h_t and w. The attention weight of each time point α_t is obtained by normalizing $h_t^{\mathsf{T}} w$ with a softmax function and sequential vector v is a weighted summation over the whole time points with attention weights.

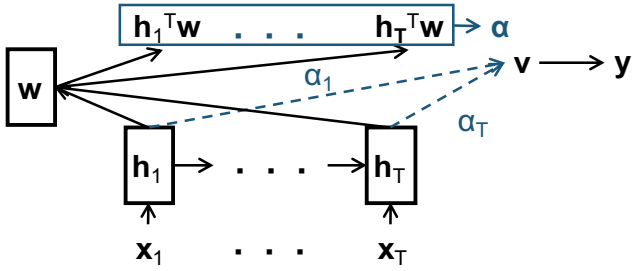

Fig. 1. Attention-based RNN for predicting labels, where x_i is an input vector

B. EEG Data Acquisition

We evaluated the methods using EEG while comprehending sentences with sequential attention models. We conducted experiments in which participants listened to sentences, some of which included semantic anomalies, and recorded the EEG signals from the participants.

1) Materials: We used previously constructed experimental materials [6] and prepared two types of language anomalies in Japanese: one is anomalies of selectional restrictions as semantic anomalies to explicitly induce incongruities and the other is syntactic violations. In this paper, we focused on analyzing the semantic conditions for classifications. The sentences were manually created [13] so that the numbers of the semantically correct sentences and incorrect sentences are identical. Table I shows an example of such paired sentences:

TABLE I
EXAMPLE OF PAIRED SEMANTICALLY CORRECT AND INCORRECT SENTENCES

a.	Taro-ga	ryoko-ni	dekake-ta
	Taro-NOM	ryoko-DAT	dekake-PAST
	(Taro set out on a journey.)		
b.	#Taro-ga	jisho-ni	dekake-ta
	Taro-NOM	jisho-DAT	dekake-PAST
	(#Taro set out on a dictionary.)		

NOM: nominative case marker;
DAT: dative case marker;
PAST: past tense morpheme.

The sentence b. that begins with # is semantically incorrect. In the experiment, each participant listened to 200 sentences: 40 semantically correct, 40 semantically incorrect, 40 syntactically correct, 40 syntactically incorrect, and 40 filler sentences.

2) Participants and Experimental Procedure: We carried out the experiment in accordance with the recommendations of ethics committee of the Nara Institute of Science and Technology. All participants wrote informed consents in accordance with the informed consents. Nineteen native Japanese speaking graduate students (16 males and 3 females) between 22 and 41 years of age (mean: 24.2) participated in this experiment. The EEG recording experiments were carried out in a soundproof room. The participants were instructed to gaze at a fixation cross displayed at the center of the monitor and avoid blinking and moving during the stimulus. The following was the experimental procedure: (1) the participants looked at the fixation cross for 1 s; (2) they listened through earphones to a randomly selected sentence for 4 s (they did not move to respond); (3) they answered whether the sentence was correct by pressing a button within 2 s. Fig.2 illustrates this procedure. All the steps for each participant were completed within 25 minutes.

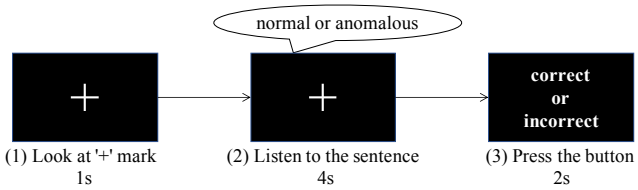

Fig. 2. Experimental design of EEG recording

3) EEG Signal Processing: We used Acticap from Brain Products with 32-ch active electrodes as an EEG cap and BrainAmp DC from the same company as an amplifier. The recorded EEG signals were processed with EEGLAB [14] to obtain relatively clean signals in the following manner. (1) Since we applied re-reference at the averaged amplitude of the TP9 and TP10 electrodes, the number of channels for the following analysis was 31. (2) The signals were passed through FIR high-pass filtering at 1 Hz. (3) We cut EEG

signals into epochs in following two conditions; i) Whole-sentence: the time regions playing whole parts of each sentence, ii) Terminal-phrase: the time regions playing only the terminal phrase of each sentence. At the same time, baseline removal was performed with the averaged amplitudes over from -500 ms to the starting point of each sentence. (4) The signals were downsampled to 250 Hz. (6) We applied ADJUST [15] to remove the components contaminated by blinking or eye movements with an independent component analysis. During the above procedure, we rejected the data of two participants' data due to a large number of rejected epochs and data defects; 1.8% of remaining data of the 17 participants was rejected.

C. Feature Extractions

We used the amplitudes at 31 channels low-pass filtered at 20 Hz for removal of noisy high frequency signals as feature vectors of the EEG signals. Therefore, the size of the vectors at each time point was 31 dimensions. Recent studies showed that neural network models have the capacity to utilize raw EEG signals as inputs by skipping some specific feature extractions [16]. We used one epoch as one datum; the temporal length of the input feature vectors corresponded to i) each spoken sentence (Whole-sentence) or ii) each terminal phrase (Terminal-phrase).

D. Training and Testing

The training data were comprised of the concatenation of 13 participants' data in which the data of two participants were used as development data for determining the optimal points for training models. The data of four different participants were testing data that validated the performances of the models with respect to the generalization of the unseen participant data. The testing data were thought to be a sufficient number of test participants to validate the usefulness based on previous works [6], [17] in which amount of about a quarter to one-third of the training sets were used for testing sets for classifications. There were 1012, 156, and 310 epochs in the training, development, and testing datasets, respectively, where the numbers of the correct and incorrect sentences were the same; therefore, the chance level of the classification was 50%. We standardized the input vectors in all the data with the mean and standard deviation (SD) of the training dataset. Each feature vector in the training dataset has a mean of 0 and an SD of 1 due to the standardization. Feature vector x_t at time t was standardized with mean μ_{train} and SD θ_{train} in the training data. The EEG data in the training dataset were augmented for the neural network models to avoid overfitting because of the small samples of experimental data. In the same manner as a previous work [18] showed in Eq.3, we added Gaussian noise to each feature of the training data to generate augmented data as follows:

$$x_{t\_aug} = x_t + 0.1 \cdot \mathcal{N}(0,1) \tag{3}$$

As an ARNN in this paper, we trained a one layered bidirectional GRU with an attention mechanism (GRU w/ att.). For the optimization of the model's hyper-parameters, we did 10-hold cross validations and found the best hyper-parameters to evaluate the model on the testing data. We empirically determined the following hyper-parameters in this order: hidden layer dimensions (5, 10, 20), augmented multiples (5, 10, 20), and the L2 regularizer weights (0, 0.1, 0.001, 0.0001). To validate the effectiveness of the attention mechanism, we compared a GRU's performance without an attention mechanism (GRU w/o att.), which had the same architecture of the model, with the best hyper-parameters except for the attention layer.

III. Results

A. Classification

Table II shows the accuracy, recall, and precision values for detecting the semantic anomalies for a bidirectional GRU with an attention mechanism (GRU w/ att.) and a bidirectional GRU (GRU w/o att.) using EEG data of region of whole sentences and region of terminal phrases. A GRU with an attention mechanism achieved 63.5% classification accuracy which was statistically and significantly higher than the chance level (two-tailed binomial test: $p < 0.01$). Fig.3 represents the accuracies of each model for each participant. This model also outperformed both of a model without attention mechanism and models with EEG of terminal phrases.

TABLE II
ACCURACY, RECALL AND PRECISION PER MODEL AND EEG DATA

Model	EEG region	Accuracy	Recall	Precision
GRU w/ att.	Whole-sentence	**0.635**	**0.716**	**0.616**
GRU w/o att.	Whole-sentence	0.554	0.470	0.565
GRU w/ att.	Terminal-phrase	0.509	0.677	0.479
GRU w/o att.	Terminal-phrase	0.467	0.516	0.470

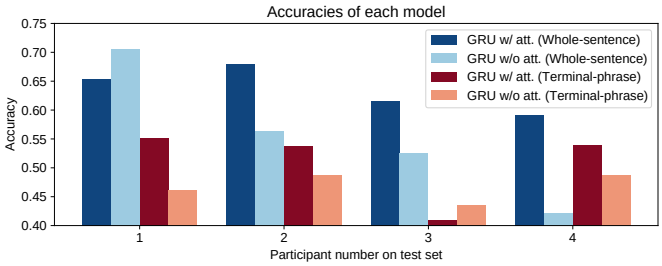

Fig. 3. Classification accuracies of each test participant

B. Visualization of Attention Weights

Visualizing the model's attention weights clarified where in the speech it depended on for the predictions. Fig.4 shows examples of the attention weights in the successful case of classifications, which demonstrate that the attention weights for predicting semantically correct sentences differed from ones for predicting semantically incorrect sentences. For predicting semantic incorrectness, the attention weights focused on the time regions close but not restricted to the onset of anomalous words (the red broken line in the left figure).

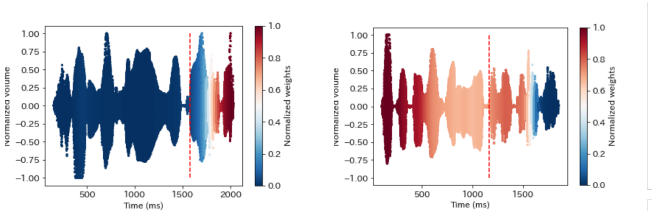

Fig. 4. Attention weights plotted on spoken sentences (left: predicting incorrectness, right: predicting correctness)

IV. Discussions

We evaluated the methods using the EEG signals of the whole areas of each sentence for detecting semantically anomalous spoken sentences and showed that the sequentially attention-based model performed well with a statistical significance level above chance and also better than the model of the previous study [6] for the same purpose. Perhaps the model predicted by identifying the relationships between the signals before and after an anomalous word with signals of the whole length of each sentence. This result demonstrated that we used more information for predictions with the signals of a sentence's whole length than of a word at a specific position. The comparison between the model with an attention mechanism and without it implied that the sequential attention weights for the EEG signals of the whole length of the sentences were feasible to classify the sequential EEG signals.

Visualizing the attention weights showed that the predictions of the models depended on the patterns of the attention weights. Therefore, the probabilities of the predictions of the semantic anomalies increased when the attention weights focused on the features temporally close to the onset of the last words in the sentences, which implied that the model learned that the signals of these time regions were important.

We investigated relationships between predicted accuracies and features of sentences. As a result, it turned out that the predicted accuracy by the attention-based model has no significant correlations to cloze probability obtained from clowdsourcing workers in [5] and length of the utterance.

V. Conclusions

We proposed a method using the EEG signals of the whole length of sentences with attention models for detecting semantically anomalous spoken sentences. Using the EEG signal data of 17 participants, the attention-based model achieved 63.5% classification accuracy with the features of the raw EEG signals that skip specific feature extractions. This result shows that the features of the whole length of the sentences were feasible for the classifications of the EEG signals and the attention mechanism worked for the sequential feature extractions for the predictions.

Future works will investigate our system's performances on sentences, including various word lengths. We will also compare performances with other feature extraction methods such as time-frequency features. In addition, experiments in other languages will show its efficiency more clearly regardless of languages. Predicting cloze probabilities, which reflect the restriction of subsequent words [4] of sentences, is another subsequent step of our work.

References

[1] A. Bakarov, "A survey of word embeddings evaluation methods," *CoRR*, vol. abs/1801.09536, 2018.

[2] S. J. Luck, *An Introduction to the Event-Related Potential Technique*. MIT Press, 2014.

[3] H.-S. Chiang, R. A. Mudar, A. Pudhiyidath, J. S. Spence, K. B. Womack, C. M. Cullum, J. A. Tanner, J. Eroh, M. A. Kraut, and J. Hart Jr, "Altered neural activity during semantic object memory retrieval in amnestic mild cognitive impairment as measured by event-related potentials," *Journal of Alzheimer's disease*, vol. 46, no. 3, pp. 703–717, 2015.

[4] M. Kutas and S. A. Hillyard, "Brain potentials during reading reflect word expectancy and semantic association," *Nature*, vol. 307, no. 5947, p. 161, 1984.

[5] S. Motomura, H. Tanaka, and S. Nakamura, "Detecting syntactic violations from single-trial eeg using recurrent neural networks," in *Adjunct of the 2019 International Conference on Multimodal Interaction*, ICMI '19, pp. 4:1–4:5, 2019.

[6] H. Tanaka, H. Watanabe, H. Maki, S. Sakriani, and S. Nakamura, "Electroencephalogram-based single-trial detection of language expectation violations in listening to speech," *Frontiers in Computational Neuroscience*, vol. 13, p. 15, 2019.

[7] Y. Roy, H. J. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *CoRR*, vol. abs/1901.05498, 2019.

[8] Z. Ni, A. C. Yuksel, X. Ni, M. I. Mandel, and L. Xie, "Confused or not confused?: Disentangling brain activity from eeg data using bidirectional lstm recurrent neural networks," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 241–246, ACM, 2017.

[9] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "Automatic sleep stage classification using single-channel eeg: Learning sequential features with attention-based recurrent neural networks," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1452–1455, IEEE, 2018.

[10] X. Zhang, L. Yao, S. S. Kanhere, Y. Liu, T. Gu, and K. Chen, "Mindid: Person identification from brain waves through attention-based recurrent neural network," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, pp. 149:1–149:23, Sept. 2018.

[11] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[12] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," *arXiv preprint arXiv:1708.00524*, 2017.

[13] S. Takazawa, N. Takahashi, K. Nakagome, O. Kanno, H. Hagiwara, H. Nakajima, and et al., "Early components of event-related potentials related to semantic and syntactic processes in the japanese language," *Brain Topography*, vol. 14, pp. 169–177, 2002.

[14] A. Delorme and S. Makeig, "Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis," *Journal of neuroscience methods*, vol. 134, no. 1, pp. 9–21, 2004.

[15] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features.," *Psychophysiology*, vol. 48 2, pp. 229–40, 2011.

[16] I. Sturm, S. Lapuschkin, W. Samek, and K.-R. Müller, "Interpretable deep neural networks for single-trial eeg classification," *Journal of Neuroscience Methods*, vol. 274, pp. 141 – 145, 2016.

[17] A. K. Vail, E. Liebson, J. T. Baker, and L.-P. Morency, "Toward objective, multifaceted characterization of psychotic disorders: Lexical, structural, and disfluency markers of spoken language," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, pp. 170–178, ACM, 2018.

[18] F. Wang, S.-h. Zhong, J. Peng, J. Jiang, and Y. Liu, "Data augmentation for eeg-based emotion recognition with deep convolutional neural networks," in *MultiMedia Modeling*, (Cham), pp. 82–93, Springer International Publishing, 2018.