LEARNING LEXICAL COHERENCE REPRESENTATION USING LSTM FORGET GATE FOR CHILDREN WITH AUTISM SPECTRUM DISORDER DURING STORY-TELLING

Yu-Shuo Liu[1]*, Chin-Po Chen*[1]*, Susan Shur-Fen Gau*[2]*, Chi-Chun Lee*[1]

[1]Department of Electrical Engineering, National Tsing Hua University, Taiwan
[1]Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taiwan
[1]cclee@ee.nthu.edu.tw, [2]gaushufe@ntu.edu.tw

ABSTRACT

Inability to carry out cohesive narratives has been identified in children with autism spectrum disorder (ASD). However, deriving cohesion measures is often done using manual labeling or relying on expert-crafted features. In this work, we develop a novel LSTM framework to learn the embedded narrative cohesion representation from data directly. Our lexical coherence representation achieves a promising recognition accuracy of 92% in classifying between typically-developing (TD) and ASD children, as compared to 73% by using conventional coherence measures computed from syntactic, word usage, and latent semantic analysis. We perform additional validity analyses on our proposed representation. By experimentally introducing incoherence in the TD's story-telling narratives through word and sentence-level shuffling, the derived lexical coherence representation from these incoherent TD data samples result in a representation closer to those of ASD data samples.

***Index Terms*—** behavioral signal processing (BSP), lexical coherence, long-short term memory neural network (LSTM), autism spectrum disorder (ASD), story-telling

1. INTRODUCTION

Autism spectrum disorder (ASD) is a highly-prevalent neurodevelopmental disorder. The perplexity of the behavior manifestations is extremely heterogeneous, which further complicates the diagnosis of ASD. Hence, developing computational framework using objective cues in supporting the diagnosis of ASD has become crucial, and it continues to be an important application domain in the emerging field of behavioral signal processing [1, 2]. Past research effort has examined the use of various behavior modalities to analyze the difference between children with autism spectrum disorder (ASD) and typically-developing (TD). For example, images of faces have been used to perform classification between ASD and TD [3], and similar tasks have been carried out in analyzing acoustic channels [4, 5] and lexical content [6, 7]. In this work, we propose a framework for deriving lexical coherence representation for children's storytelling narration and use it in the task of differentiating between ASD and TD.

Past research in analyzing narratives of ASD children in storytelling have used discourse-related measures and several other text similarity features to identify idiosyncratic words and topics demonstrating the differences in the word usage between ASD and TD [8, 9]. In terms of lexical coherence, several studies have indicated that it is difficult for individuals with autism to tell a story in a coherent sequence [10], i.e., autistic children are less likely to include causal statements in their story narration. Instead, they tend to only elaborate the local story episodes without composing a coherent whole [11, 12]. The lack of lexical coherence has also been reported in ASD by Diehl et al. [13]. The work done by Regneri et al. further conducts context annotation to measure the narrative cohesion of children with ASD as an indicator of their narration ability [6]. Most of these methods rely heavily on expert human annotation, and the labeling is often restricted to a particular kind of narratives.

In this work, we propose a data-driven method to learn the lexical coherence representation by projecting word vectors into a long short-term memory neural network's (LSTM) forget gate, in which the embedded contextual information can be uncovered. We conduct classification experiment in differentiating between children of TD versus ASD during story-telling. We obtain an promising unweighted average recall (UAR) of 0.92 using our lexical coherence representation, which is a relative improvement of 26.02% over the well-known coherence measures proposed by McNamara [14, 15]. Furthermore, similar to recent works [16, 17], we experimentally modify our input data to observe the variation of LSTMs internal parameters. By shuffling either word order and/or sentence sequence in the TD's story-telling samples, this embedded representation would become less "coherent" and move closer to the representation of ASD. The rest of the paper is organized as follows: database description and the proposed framework are detailed in section 2, experimental results in section 3, and finally conclusion in section 4.

2. RESEARCH METHODOLOGY

2.1. Datasets

We utilize two different story-telling narrative corpora in this work. The first corpus (Dataset I) is a 'ASD and TD story-

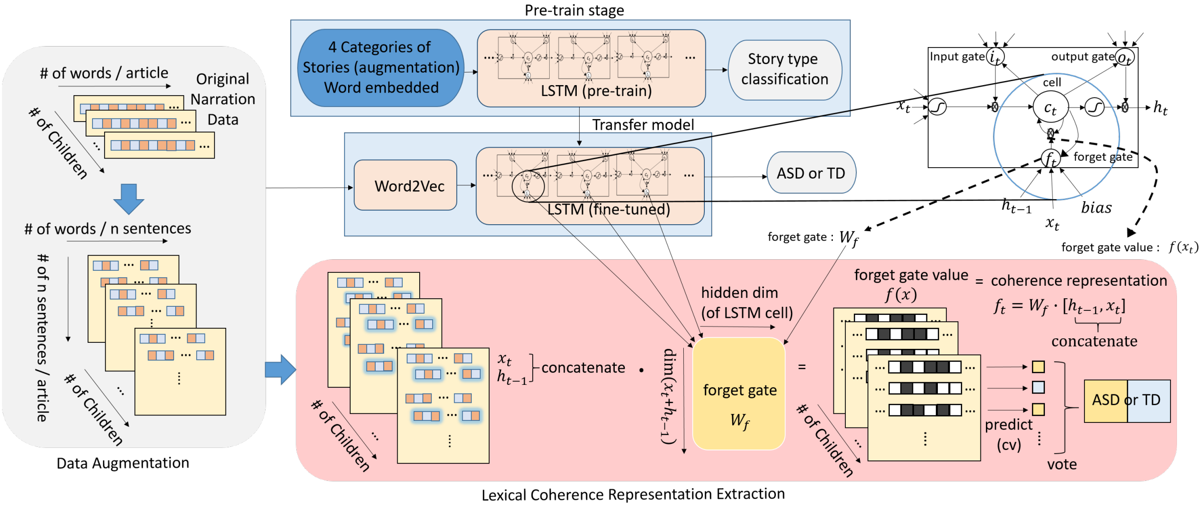


Fig. 1. It shows a complete architecture of our proposed data-driven lexical coherence representation. It includes components of: data augmentation, Chinese Word2Vec, LSTM training on Dataset II, fine-tuning on Dataset I, and finally extraction of coherence representation from LSTM forget gate.

Table 1. Informations of participants in Dataset I

Story-telling corpus	ASD (31)	TD (36)
Age (Avg/Std)	15.3/3.72	12.4/0.80
Words/Story (Avg/Std)	465.03/267.98	385.29/169.77

telling' database, which we use to examine the discriminative power of our data-driven lexical coherence representation. The second corpus (Dataset II) is a background 'fairy-tale' story-telling corpus, which we use to construct the background LSTM.

2.1.1. *Dataset I: ASD and TD Story-telling*

Dataset I is a corpus collected by asking children to construct a story with the picture book, **'The Tuesday Story'**. Each page of the book is accompanied by a sentence and the main picture illustration.

For the ASD participants, this storytelling is collected as part of the instrumentation of the Autism Diagnostic Observational Schedule (ADOS) [18], which is administered at the Department of Psychiatry, National Taiwan University Hospital (NTUH)[1]. ADOS is a gold-standard clinically-valid instrument in eliciting natural and targeted social communicative behaviors of the participant through semi-structured dyadic interactions. For the TD subjects, we follow the exact dyadic interview-style of ADOS interactions closely, i.e., instructing the subjects to narrate the same picture book. As an example, the interviewee would start the story-telling session:

"The story begins on Tuesday evening. A group of frogs start their journey at some wetlands, then fly to the nearest town...now you continue the story..."

The ASD and TD Story-telling dataset consists a total of 67 subjects with approximately 28,446 Chinese words (average number of words of narration is 424.56 words). The informations of the participants is presented in Table 1.

Table 2. Story type distribution in Dataset II

Fairy-tale corpus (# of article)	average number of words
idiom story(24)	624.25
bed time story(24)	571.04
contemporary fairy tale(24)	682.66
puzzle story(24)	589.20

2.1.2. *Dataset II: Fairy-tale Story*

Dataset II is a corpus crawled from online website[2]. We randomly choose 4 categories of story-telling corpus whose article lengths are similar to each other. The 4 categories are idiom story, bed time story, contemporary fairy tale, and puzzle story. We randomly choose 24 articles to constitute our dataset from each category. The total number of articles is 96 with approximately 59,212 Chinese words. Table 2 summarizes the distribution of story type in Dataset II.

2.2. Lexical Coherence Representation

Figure 1 shows a complete architecture of our proposed lexical coherence representation. It includes components of: data augmentation, Chinese Word2Vec, LSTM training on Dataset II, fine-tuning on Dataset I, and finally coherence representation extraction from LSTM forget gate.

2.2.1. *Data Augmentation & Chinese Word2Vec*

Researchers have found using data augmentation, it helps control the generalization error of neural network, e.g., in neural translation and speech recognition [19, 20]. In this paper, we use a simple sliding window approach to perform data augmentation on both datasets. We take **n** sentences as a data

[1] Approved by IRB: REC-10501HE002 and RINC-20140319

[2] http://wap.etgushi.com/index.html

sample assigned with the same label as the entire document, and we shift one sentence at a time to generate another data sample with **n** sentences. The choice of **n** is done empirically from the set of $\{1, 3, 5, 7, 9\}$.

Furthermore, word2vec has become the most general neural representation of words in recent years [21]. We first perform Chinese word segmentation using Jieba toolbox [22]. The Chinese word2vec is trained on both datasets using continuous bag-of-word (CBOW) approach. Each Chinese word is represented by a 32-dimensional word vector in our work.

2.2.2. *Long Short-Term Memory (LSTM) Neural Network*

Our coherence representation requires extracting the output of forget gate after projecting narrative samples into an LSTM. In this section, we will first briefly describe LSTM. An LSTM is a directional time-series neural network [23]. The core of LSTM is the information contained in the cell state $\widetilde{C_t}$ that is updated at every time step:

$$\widetilde{C_t} = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right)$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}$$

The benefit of using LSTM modeling is in its ability to regulate the amount of information retained in the long or short term memory context. The regulation mechanism is done using the structure of gates, which is formulated as a weight and a sigmoid layer. Each LSTM has three gates (f_t, i_t, o_t):

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right)$$

In specifics, the forget gate, f_t, is the gate responsible for controlling the amount of past information being let through into the main cell $\widetilde{C_t}$, which we leverage as our main mechanism in deriving the lexical coherence measure. We first train a supervised background LSTM$_{\text{story}}$ with 128 hidden dimensions on Dataset II on labels of the four story types.

2.2.3. *Lexical Coherence Representation*

In order to derive the lexical coherence representation, we first perform fine-tuning of LSTM$_{\text{story}}$ on Dataset I to obtain LSTM$_{\text{TD-ASD}}$ through weight-sharing. Then, by using the learned weight matrix, W_f, for the forget gate, we can get the computed output value, $\overline{f_t}$, just prior to the activation function without the bias term:

$$\overline{f_t} = W_f \cdot [h_{t-1}, x_t]$$

Hence, for every data sample **t**, i.e., **n** sentences where each sentence has **k** words, there would be a total $\mathbf{n} \times \mathbf{k}$ number of $\overline{f_t}$ sequence. We encode them to a fixed representation for every sample **t**,

$$F = g\left(\overline{f_1}, ..., \overline{f_T}\right)$$

where g indicates 17 statistical functionals, (max, min, mean, median, standard deviation, 1st percentile, 99th percentile, 99th percentile - 1st percentile, skewness, kurtosis, minmum position, maximum position, lower quartile, upper quartile, interquartile range, power, and 1st difference). This F is our lexical coherence representation for each data sample.

Table 3. Summary of TD vs. ASD Classification Results

Features	UAR
Coh-Metrix	0.73
TFIDF	0.77
TFIDF + Coh-Metrix	0.80
LSTM	0.85
Lexical Coherence Representation	**0.92**

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup

In this work, we use our proposed lexical coherence representation to perform classification between TD and ASD. We compare to the following methods:

- Term-Frequency Inverse Document-Frequency (**TFIDF**): A 2034 dimensional feature vector in this work.
- **Coh-Metrix**: Coh-Metrix is considered as one of the most sophisticated automated evaluation of text and discourse. It computes coherence metrics of written and spoken texts [15]. The measures include similarity computed between neighboring sentences, nouns overlap (repetition) in neighboring sentences or an entire article, and the similarity of sentences structure of an article, etc. We use all 20 dimensional measurement[3].
- **LSTM**: Training a TD vs. ASD, supervised LSTM directly and pull the last time step's hidden layer just prior to the softmax layer as features.
- **Lexical Coherence Representation**: Using our proposed coherence presentation as features.

The classifier used is support vector machine, and correlation based feature selection is also carried out. Since every narrative includes a different number of data samples, in order to come up with a single decision at the narrative-level for an individual subject, we perform major vote over the prediction results. The evaluation scheme is carried out using leave-one-subject-out cross validation, and the metric used is the unweighted average recall (UAR).

3.2. Experimental Results

First of all, the feature selection results from the TFIDF methods show that conjunction, adverb and common catch phrase in Chinese such as ”and then,” ”so,” ”that,” ”like this” are significantly higher in the narration of ASD than in TD subjects. Several key-words about the story content, such as “rooftop,” “old grandma,” “floor,” “live in”, on the contrary,

[3] http://cohmetrix.com/

Table 4. Comparing UAR of lexical representation with different choices of **n**-sentence as a data sample

# of sentence	LSTM	Lexical Coherence
n=1	0.76	0.78
n=3	0.85	0.87
n=5	0.78	0.92
n=7	0.79	0.81
n=9	0.71	0.81

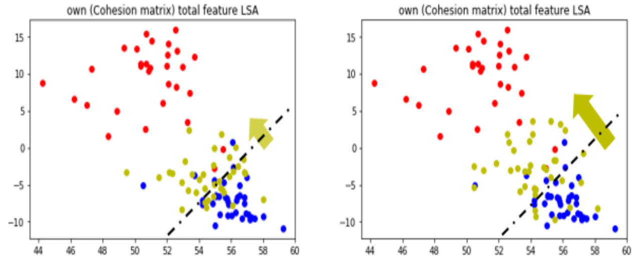
are significantly higher in TD subjects. Furthermore, the most important features from coh-metrix are "nouns repetition of the neighboring sentence" and "nouns repetition of the whole article", "sentence similarity of neighboring sentences", and "sentence similarity of whole article". Many of these findings are indeed intuitively satisfying.

In terms of classification accuracy, Table 3 summarizes our experimental results. The best accuracy obtained is by using our proposed lexical coherence measures, which achieves an UAR of 92%. It outperforms both the baseline coh-metrix and TDIDF method, 73% and 77%, by a significant margin. Also, it is interesting to see that when comparing to using LSTM directly to perform classification, the use of our proposed coherence measure outperforms by 8% absolute. In the context of differentiating between TD vs. ASD, the measure of lexical coherence, which is derived from an internal parameter of an LSTM, seems to be more indicative than using the entire LSTM. Finally, the effect of the number of sentences, **n**, that forms a data sample in obtaining the lexical coherence representation on the accuracies obtained is listed in Table 4. The optimal number of sentences seems to be around lexical coherence is 5.

3.3. Analysis of Lexical Coherence Representation

Our proposed method achieves high TD vs. ASD recognition accuracy. Due to the complexity of the model, it is difficult to interpret our feature representation directly. In this work, we adopt a similar approach recently published [16, 17] to understand our framework, i.e., by experimentally creating various realizations of our intended construct (lexical coherence) within our dataset to examine the change in our derived representation. In this analysis, we first create a *word incoherency* by randomly shuffling the word order in the TD's data samples of Dataset I. We can then visualize our derived lexical representation in a 2-D plot using latent semantic analysis (LSA) based projection (Figure 2a). Then we further randomly reshuffle the sentence order to create additional *sentence incoherency*. This is again plotted in the 2-D projection as shown in Figure 2b.

The BLUE dots indicate the data samples of TD subjects, the RED dots denotes ASD subjects, and the YELLOW dots represent our simulated data samples by shuffling either word or word+sentence order. It is interesting to observe that in Figure 2a, the YELLOW dots are sitting right in between TD and ASD samples. Furthermore, as we introduce more incoherency (from *word incoherency* to *word incoherency + sentence incoherency*) into the TD samples, the YELLOW dots move even closer to the ASD samples. This experiment provides evidence in showing that our proposed representation reflect the intended construct to measure - the lexical coherence in story-telling.



(a) random word order (b) random sentence order

Fig. 2. a) word incoherency LSA projection and b) sentence incoherency LSA projection. Blue: TD, Red: ASD, Yellow: Simulated incoherent data

4. CONCLUSION AND FUTURE WORK

In this paper, we have presented a novel data-driven lexical coherence feature representation learned by using LSTM network, especially leveraging the forget gate. The derived representation provides 92% recognition accuracy in differentiating between TD vs. ASD on narratives of story-telling. It has also been shown to outperform conventional measures of lexical content and coherence. Lastly, by simulating incoherent structure in the TD's narratives, we can visually observe that as we introduce more incoherent variations, our derived lexical representation of TD samples move closer to that of ASD samples. It is exciting to see that while the representation does not explicitly learn on labels of coherency, it seems to capture such a construct within the internal parameters of LSTM; at the same time, it achieves a high classification rate.

There are multiple future directions to pursue. One of the immediate work is to examine the relationship between this data-derived lexical coherence representation directly with semantically-meaningful coherence measures previously proposed in the computational linguistics for a detailed understanding of this model. Furthermore, the construct of coherency can be abstracted into speech acoustic domain (e.g., fluency in intonation) and gestural dynamics (e.g., coordinative aspect of hand gestures and head movement), we will explore the multimodal aspect of coherence features. Lastly, realizing this behavior informatics in the real world setting of clinical value will continue to be a central goal.

5. REFERENCES

[1] Shrikanth Narayanan and Panayiotis G Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[2] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan, "Signal processing and machine learning for mental health research and clinical applications [perspectives]," *IEEE Signal Processing Magazine*, vol. 34, no. 5, pp. 196–195, 2017.

[3] Wenbo Liu, Ming Li, and Li Yi, "Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework," *Autism Research*, vol. 9, no. 8, pp. 888–898, 2016.

[4] Erik Marchi, Björn Schuller, Simon Baron-Cohen, Ofer Golan, Sven Bölte, Prerna Arora, and Reinhold Häb-Umbach, "Typicality and emotion in the voice of children with autism spectrum condition: Evidence across three languages," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Daniel Bone, Chi-Chun Lee, Matthew P Black, Marian E Williams, Sungbok Lee, Pat Levitt, and Shrikanth Narayanan, "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *Journal of Speech, Language, and Hearing Research*, vol. 57, no. 4, pp. 1162–1177, 2014.

[6] Michaela Regneri and Diane King, "Automated discourse analysis of narrations by adolescents with autistic spectrum disorder," *ACL 2016*, p. 1, 2016.

[7] Arodami Chorianopoulou, Efthymios Tzinis, Elias Iosif, Asimenia Papoulidi, Christina Papailiou, and Alexandros Potamianos, "Engagement detection for children with autism spectrum disorder," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5055–5059.

[8] Masoud Rouhizadeh, Emily Prud'Hommeaux, Brian Roark, and Jan Van Santen, "Distributional semantic models for the evaluation of disordered language," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. NIH Public Access, 2013, vol. 2013, p. 709.

[9] Masoud Rouhizadeh, Richard Sproat, and Jan Van Santen, "Similarity measures for quantifying restrictive and repetitive behavior in conversations of autistic children," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. NIH Public Access, 2015, vol. 2015, p. 117.

[10] Molly Losh and Peter C Gordon, "Quantifying narrative ability in autism spectrum disorder: A computational linguistic analysis of narrative coherence," *Journal of autism and developmental disorders*, vol. 44, no. 12, pp. 3016–3025, 2014.

[11] Lisa Capps, Molly Losh, and Christopher Thurber, "the frog ate the bug and made his mouth sad: Narrative competence in children with autism," *Journal of abnormal child psychology*, vol. 28, no. 2, pp. 193–204, 2000.

[12] Helen Tager-Flusberg, "once upon a ribbit: Stories narrated by autistic children," *British journal of developmental psychology*, vol. 13, no. 1, pp. 45–59, 1995.

[13] Joshua J Diehl, Loisa Bennetto, and Edna Carter Young, "Story recall and narrative coherence of high-functioning children with autism spectrum disorders," *Journal of abnormal child psychology*, vol. 34, no. 1, pp. 83–98, 2006.

[14] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior Research Methods*, vol. 36, no. 2, pp. 193–202, 2004.

[15] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai, *Automated evaluation of text and discourse with Coh-Metrix*, Cambridge University Press, 2014.

[16] Jiwei Li, Will Monroe, and Dan Jurafsky, "Understanding neural networks through representation erasure," *arXiv preprint arXiv:1612.08220*, 2016.

[17] Pang Wei Koh and Percy Liang, "Understanding black-box predictions via influence functions," *arXiv preprint arXiv:1703.04730*, 2017.

[18] Catherine Lord, Michael Rutter, Pamela C.. Dilavore, and Susan Risi, *ADOS: Autism diagnostic observation schedule*, Hogrefe Boston, 2008.

[19] Marzieh Fadaee, Arianna Bisazza, and Christof Monz, "Data augmentation for low-resource neural machine translation," *arXiv preprint arXiv:1705.00440*, 2017.

[20] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[22] J Sun, "jiebachinese word segmentation tool," 2012.

[23] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.