Topic-Based Measures of Conversation for Detecting Mild Cognitive Impairment

Liu Chen
Center for Spoken Language Understanding
Oregon Health & Science University
chliu@ohsu.edu

Hiroko H Dodge
Department of Neurology
Oregon Health & Science University
dodgeh@ohsu.edu

Meysam Asgari
Center for Spoken Language Understanding
Oregon Health & Science University
asgari@ohsu.edu

Abstract

Conversation is a complex cognitive task that engages multiple aspects of cognitive functions to remember the discussed topics, monitor the semantic and linguistic elements, and recognize others' emotions. In this paper, we propose a computational method based on the lexical coherence of consecutive utterances to quantify topical variations in semi-structured conversations of older adults with cognitive impairments. Extracting the lexical knowledge of conversational utterances, our method generates a set of novel conversational measures that indicate underlying cognitive deficits among subjects with mild cognitive impairment (MCI). Our preliminary results verify the utility of the proposed conversation-based measures in distinguishing MCI from healthy controls.

1 Introduction

Speech and language characteristics are known to be effective social behavioral markers that could potentially serve to facilitate the identification of measured "markers" reflecting early cognitive changes in at-risk older adults. Recent advances on natural language processing (NLP) algorithms have given the researchers the opportunity to explore subtleties of spoken language samples and extract a wider range of clinically useful measures. Leveraging an NLP-based method, our objective in this study is to characterize the ongoing dynamics of topics over the course of everyday conversation between an interviewer and an older adult with or without cognitive impairment. Our proposed method translates its analysis of conversation into a set of quantifiable measures that can be used in clinical trials for early detection of a cognitive deficit. Our cohort includes a professionally transcribed dataset of 30-minute audio recordings collected from conversation-based social interactions carried out between standardized interviewers and participants with either normal cognition or MCI (clinicaltirals.gov: NCT02871921). We evaluate the utility of proposed conversation-based measures in detecting MCI incidence. To the best of our knowledge, analysis of exchanged topics in conversations have not been used to examine the cognitive status of older adults.

1.1 Conversational Speech and Cognitive Impairment

Recent studies have attempted to leverage natural language processing (NLP) algorithms to automatically characterize atypical language characteristics observed in age-related cognitive decline(Roark et al., 2011; Asgari et al., 2017; Shibata et al., 2016; Mueller et al., 2016). With a few exceptions, most of these studies have used elicited speech paradigms to generate speech samples, for example, using traditional neuropsychological language tests such as the verbal fluency test (citing names from a semantic category such as animals or fruits within a short amount of time) or the story recall test (recalling specific stories subjects are exposed to during a testing session). As a result, their assessment of language characteristics is constrained by the nature of language tests. Alternatively, everyday conversations have been recently explored to gain insight about the consequences of a cognitive deficit on a patient's speech and language characteristics (Khodabakhsh et al., 2015; López-de Ipina et al., 2015; Hoffmann et al., 2010). Semi-structured conversations (i.e., talk about pre-specified topics) more closely resemble to naturalistic speech than elicited speech tasks (e.g., verbal fluency tests, picture naming tests) and provide a rich source of information allowing us to correlate various aspects of spoken language to cognitive functioning. Conversation is a complex cognitive task that engages multiple domains of

Proceedings of the 1st Workshop on NLP for Medical Conversations, pages 63–67
Online, 10, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

cognitive functions including executive functions, attention, working memory, memory, and inhibition to control the train of thoughts, and to monitor semantic and linguistic elements of the discourse. It also involves social cognition to understand others' intentions and feelings (Ybarra, 2012; Ybarra et al., 2008). Quantifying atypical topic variations in prodromal Alzheimer's disease represents an important, and yet under-examined area that may reveal underlying cognitive processes of patients with MCI.

1.2 Topic Segmentation

A key problem in our conversation analysis is dividing the consecutive utterances into segments that are topically coherent. This is a prerequisite step for our higher-level analysis of conversations involving representation of entire conversation by a set of quantifiable measures. Topic segmentation methods first segment the sequence of utterances into a set of finite topics, representing utterances as vectors in a semantic space. Next, they measure the correlation between two adjacent encoded utterances, and finally predict the topic boundary according to a pre-specified threshold value compared to calculated correlations. Based upon the criteria they adopt for quantifying the cohesion among a pair of consecutive utterances, they can be broadly categorized into two models. Assuming the topic shifting is strongly correlated to the term shifting, *lexicon cohesion models* rely on similar terms of each utterance; that is, topically coherent utterances share some common terms within a short window of spoken words. They are learned in an unsupervised fashion and do not require labeled data. Widely used algorithms such as *TextTiling* (Hearst, 1997) and *LCSeg* (Galley et al., 2003) are examples of lexical based methods for topic segmentation. In contrast to lexical based methods, *contextual cohesion models* exploit the semantic knowledge from the entire utterance rather than key terms. These context-dependent models assume that utterances with a similar semantic distribution share the same topic. More recent methods leverage the deep architectures, such as recurrent neural networks (RNNs) (Sehikh et al., 2017) and convolutional neural networks (CNNs) (Wang et al., 2016) to semantically encode the utterance into a vector space. Treating the topic segmentation as a sequence labeling problem, labels (i.e., topics) are then assigned to every utterance. Context dependent models assume that, if two documents share the same topic, the word distribution of these two should also be similar. Despite the potential benefits of extracting the knowledge from the content, there exist several barriers to taking advantage of them in clinical conversations. Successful deep architectures are trained on large amounts of training examples, typically obtained from structured written text such as medical textbooks or Wikipedia. These models perform well in highly structured data; however, their performance degrades once used in unstructured samples, such as social conversations, due to mismatch between the characteristics of testing and training examples. Topic segmentation in conversational text is more challenging than the written text as it is less structured and typically include shorter utterances (e.g., acknowledgements) and disfluencies (e.g., "um" and "hmm").

2 Data collection and participants

For this preliminary work, we used a collection of semi-structured conversations collected randomized controlled clinical trial entitled *I-CONECT* (https://www.i-conect.org/; ClincialTrials.gov: NCT02871921) conducted at Oregon health Science University (OHSU), University of Michigan, and Wayne State University. In *I-CONECT* study, participants engage in a 30-minute video chat 4 times per week for 6 months (experiment group) followed by 2 times per week for an additional 6 months (control group). Conversations are semi-structured, in which participants freely talk about a predefined topic such as leisure time, science, etc. with trained interviewers. Interviewers were asked to engage participants into a conversation by showing picture prompts, share facts, and ask questions related to predefined topics such as leisure time and science. Interviewers were also instructed to minimally contribute to the conversation (less than 30% of total conversation time) and let participants freely talk about daily selected topics. Our analysis includes a total of 45 older adults, 23 with MCI and 22 healthy controls. Table 1 reports their baseline characteristics. Upon completion of Montreal Cognitive Assessment (MoCA) (Nasreddine et al., 2005), a cognitive screening tool to identify MCI, the test results were evaluated at consensus meeting to clinically determine MCI or normal (i.e., clinicians' consensus based-determination).

Variable	Intact n=22	MCI n=23
Age	80.82 (4.87)	84.06 (5.43)
Gender (% Women)	86.36%	68.22%
Years of Education	16.05 (2.70)	15.17 (2.85)
MoCA	26.14 (2.46)	22.00 (2.84)

Table 1: Baseline characteristics of MCI and cognitively intact participants. Montreal Cognitive Assessment (MoCA) score, ranged from 0 to 30, is used as a screening tool and it is lower in MCI subjects.

3 Methods

In our recent study, we presented a method for automatically identifying individuals with MCI based on the count of individuals' spoken words taken from the semi-structured conversations between interviewers and participating older adults (H Dodge et al., 2015; Asgari et al., 2017). We showed that individuals with MCI talk more than healthy controls in these conversations (H Dodge et al., 2015), as they may need to substitute words in the conversation to convey their thoughts. Also, we showed that their lexical pattern, obtained by counting the frequency of words picked from a particular word category such as *verbs* and *fillers*, is different from healthy controls (Asgari et al., 2017). The main limitation of our prior works on linguistic analysis of conversations is ignoring sentence structure and other contextual information relying entirely on word-level features. Enhancing our automatic analysis of clinical conversation, we aim to characterize the relationship among the sequence of sentences, presented in the course of conversation, in order to track the exchanged topics. Our central hypothesis in this work is that patients with MCI may have subtle difficulties with executive and self-monitoring conversation consistency relative to those with normal cognition resulting in more disruptive pattern of exchanged topics within the conversation.

3.1 Utterance Representation

Given the limited amounts of text data in this study, it is difficult to employ deep architectures for learning semantic models. Instead, we adopt LCseg (Galley et al., 2003) algorithm to divide utterances into semantically related clusters. LCseg uses word repetitions to build lexical chains that are consequently used to identify and weight the key terms. A lexical chain is a set of semantically related words inside a window of utterances that capture the lexical cohesion followed within the window. From the lexical chains, it then computes lexical cohesion (LC) score among two adjacent analysis windows utterances.

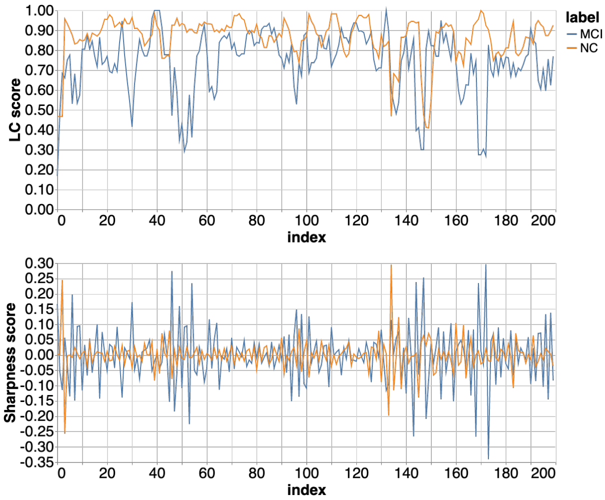

Figure 1: LC (top) and sharpness (bottom) scores of two MCI and NC subjects as a function of utterance index.

To predict a topic boundary, LCseg tracks the fluctuation of LC scores and estimates an occurrence of a topic change according to a sharpness measure calculated on surrounding left and right neighbors of the ith center window as :

$$S_i = \frac{1}{2}[LC_{i-1} + LC_{i+1} - 2 * LC_i] \qquad (1)$$

Assuming that sharp changes in sharpness score co-occur with a change in the topic, LCseg locates the topic boundaries where the sharpness score exceeds a pre-specified threshold value. LCseg was originally designed to analyze transcription of multiparty oral meetings that typically include six to eight participants. Similar to our semi-structured conversations, ungrammatical sentences are common in such meetings.

3.2 Automatic Measures of Conversation

The top plot in Figure (1) depicts the lexical cohesion scores calculated across the sequence of utterances chopped from conversation recordings of two MCI and normal control (NC) participants. The horizontal axis represents the utterance index that spans from the beginning to the end of the conversation, and the vertical axis represents the lexical cohesion score. As it is seen in these plots, the LC scores of the normal control (NC) participant are smoother with less frequent sharp changes com-

model	ROC AUC	Sensitivity	Specificity	Accuracy
SVM	83.82% (13.39%)	80.77% (19.57%)	77.36% (18.25%)	79.15% (12.44%)

Table 2: Classification results (with standard deviations) for distinguishing 23 MCI from 22 normal controls.

pared to participants with MCI, suggesting a structural difference in the pattern of their discussed topics across the conversation. To measure the variations of the LC score across the utterances, we use Shannon's entropy, an appropriate metric to measure the level of organization in random variables (Renevey and Drygajlo, 2001) and measure the entropy of harmonic coefficients. The bottom plot in Figure (1) depicts the sharpness score calculated on LC score of two MCI and NC participants (top plot) according to Equation 1. The more frequent and yet abrupt changes in sharpness score of MCI subject indicates the higher likelihood of topical changes in the sequence of utterance compare to the NC subject. To capture the frequency of these changes, we adopt the zero-crossing rate (ZCR), a measure that quantifies the number of times a signal crosses the zero line within a window of the signal. ZCR is a common measure in speech processing algorithms for differentiating speech from noise segments (Bachu et al., 2010). Prior to compute the ZCR, we normalize the sharpness score such that it becomes a zero-mean signal. Dividing the entire signal into finite number of fixed-length windows, we compute the ZCR for every window and ultimately summarize the computed ZCRs across the entire conversation using mean and summation statistical functions.

4 Experiments

4.1 Pre-processing and Feature Extraction

Removing the interviewer's speech, we narrow our focus on the analysis of the participant's side of the conversation. For pre-processing of the transcriptions (e.g, removing the punctuation), we adopt an open-source library, SpaCy (Honnibal and Montani, 2017), with its default settings. We also set the minimum number of words per utterance to three words and exclude the shorter utterances. We also trimmed out fillers (e.g., "hmm", "mm-hmm", and "you know") from the transcriptions. Pre-processed transcription of conversations are then fed into LC-seg algorithm where from its output, LC score, we compute the sharpness score. Next, we calculate the entropy of the LC score as well as ZCR of both LC score and sharpness score as described at 3.2.

4.2 Results

Representing a conversation using four measures selected by RFECV (sum and mean of ZCR on LC score, the entropy of LC score as well as the sum and mean of ZCR on sharpness scores), we trained a linear support vector machine (SVM) classifier from the open-source Scikit-learn toolkit (Pedregosa et al., 2011) to validate the utility of proposed conversation measures in distinguishing MCI from NC participants. We used cross-validation (CV) techniques in which the train and test sets are rotated over the entire data set. We shuffle the data and repeat 5-fold cross-validation 100 times. Our results, reported in Table (2), present the mean and standard deviation of four classification metrics: 1) sensitivity, 2) specificity, 3) area under the curve of receiver operating characteristics (AUC ROC), and 4) classification accuracy. Our results indicates that our proposed measures are useful in detecting subjects with MCI.

5 Conclusion

In our clinically oriented study, conversations between the interviewer and the participant provide an opportunity to analyze potential differences in the conversational output of persons with MCI and cognitively intact adults. With the aim of gaining insight about the underlying cognitive processing among patients with MCI, we proposed a computational approach to capture atypical variations observed in the sequence of topics discussed throughout the course of conversation. Our method represents the entire conversation with a set of quantifiable measures that are useful in early detection of cognitive impairment. Despite this promise, a current important limitation to this approach is that the analysis relies on high-fidelity transcription of the conversations which is labor intensive. Furthermore, when applying this approach in clinical trials or to the general population, one would typically add other potentially predictive features to the classification model such as age, gender, education, and family history of dementia. Future studies will need to examine larger and more diverse populations over time and explore the possible cognitive bases behind the findings of the present study.

Acknowledgments

This work was supported by Oregon Roybal Center for Aging and Technology Pilot Program award P30 AG008017-30 in addition to NIH-NIA Aging awards R01-AG051628, and R01-AG056102.

References

Meysam Asgari, Jeffrey Kaye, and Hiroko Dodge. 2017. Predicting mild cognitive impairment from spontaneous spoken utterances. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):219–228.

RG Bachu, S Kopparthi, B Adapa, and Buket D Barkana. 2010. Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. In *Advanced Techniques in Computing Sciences and Software Engineering*, pages 279–282. Springer.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. 2003. Discourse Segmentation of Multi-Party Conversation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 562–569. Association for Computational Linguistics.

Hiroko H Dodge, Nora Mattek, Mattie Gregor, Molly Bowman, Adriana Seelye, Oscar Ybarra, Meysam Asgari, and Jeffrey A Kaye. 2015. Social Markers of Mild Cognitive Impairment: Proportion of Word Counts in Free Conversational Speech. *Current Alzheimer Research*, 12(6):513–519.

Marti A Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational linguistics*, 23(1):33–64.

Ildikó Hoffmann, Dezso Nemeth, Cristina D Dye, Magdolna Pákáski, Tamás Irinyi, and János Kálmán. 2010. Temporal parameters of spontaneous speech in Alzheimer's disease. *International journal of speech-language pathology*, 12(1):29–34.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Karmele López-de Ipina, Jordi Solé-Casals, Harkaitz Eguiraun, Jesús B Alonso, Carlos M Travieso, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, and Blanca Beitia. 2015. Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach. *Computer Speech & Language*, 30(1):43–60.

Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. 2015. Evaluation of linguistic and prosodic features for detection of Alzheimer's disease in Turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):9.

Kimberly Diggle Mueller, Rebecca L Koscik, Lyn S Turkstra, Sarah K Riedeman, Asenath LaRue, Lindsay R Clark, Bruce Hermann, Mark A Sager, and Sterling C Johnson. 2016. Connected Language in Late Middle-Aged Adults at Risk for Alzheimer's Disease. *Journal of Alzheimer's Disease*, 54(4):1539–1550.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. 2005. The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool For Mild Cognitive Impairment. *Journal of the American Geriatrics Society*, 53(4):695–699.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Philippe Renevey and Andrzej Drygajlo. 2001. Entropy Based Voice Activity Detection in Very Noisy Conditions. In *Seventh European Conference on Speech Communication and Technology*.

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken Language Derived Measures for Detecting Mild Cognitive Impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

Imran Sehikh, Dominique Fohr, and Irina Illina. 2017. Topic segmentation in ASR transcripts using bidirectional rnns for change detection. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 512–518. IEEE.

Daisaku Shibata, Shoko Wakamiya, Ayae Kinoshita, and Eiji Aramaki. 2016. Detecting Japanese Patients with Alzheimer's Disease based on Word Category Frequencies. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 78–85.

Liang Wang, Sujian Li, Xinyan Xiao, and Yajuan Lyu. 2016. Topic Segmentation of Web Documents with Automatic Cue Phrase Identification and BLSTM-CNN. In *Natural Language Understanding and Intelligent Applications*, pages 177–188. Springer.

Oscar Ybarra. 2012. On-line Social Interactions and Executive Functions. *Frontiers in human neuroscience*, 6:75.

Oscar Ybarra, Eugene Burnstein, Piotr Winkielman, Matthew C Keller, Melvin Manis, Emily Chan, and Joel Rodriguez. 2008. Mental Exercising Through Simple Socializing: Social Interaction Promotes General Cognitive Functioning. *Personality and Social Psychology Bulletin*, 34(2):248–259.