A Digital Language Coherence Marker for Monitoring Dementia

Anonymous ACL submission

Abstract

The use of spontaneous language to derive appropriate digital markers has become an emergent, promising and non-intrusive method to diagnose and monitor dementia. Here we propose methods to capture language coherence as a cost-effective, human-interpretable digital marker for monitoring cognitive changes in people with dementia. We introduce a novel task to learn the temporal logical consistency of utterances in short transcribed narratives and investigate a range of neural approaches. We compare such language coherence patterns between people with dementia and healthy controls and conduct a longitudinal evaluation against three clinical bio-markers to investigate the reliability of our proposed digital coherence marker. The coherence marker shows robustness in distinguishing between people with mild cognitive impairment, those with Alzheimer's Disease and healthy controls. Moreover our analysis shows high association between the coherence marker and the clinical bio-markers as well as generalisability potential to other related conditions.

1 Introduction

Dementia includes a family of neurogenerative conditions that affect cognitive functions of adults. Early detection of cognitive decline could help manage underlying conditions and allow better quality of life. Many aspects of cognitive disorders manifest in the way speech is produced and in what is said (Forbes-McKay and Venneri, 2005; Voleti et al., 2019). Previous studies showed that dementia is often associated with thought disorders relating to inability to produce and sustain coherent communication (McKhann, 1987; Hoffman et al., 2020). Language coherence is a complex multifaceted concept which has been defined in different ways and to which several factors contribute (Redeker, 2000). A high-quality communication is logically consistent, topically coherent, and pragmatically reasonable (Wang et al., 2020).

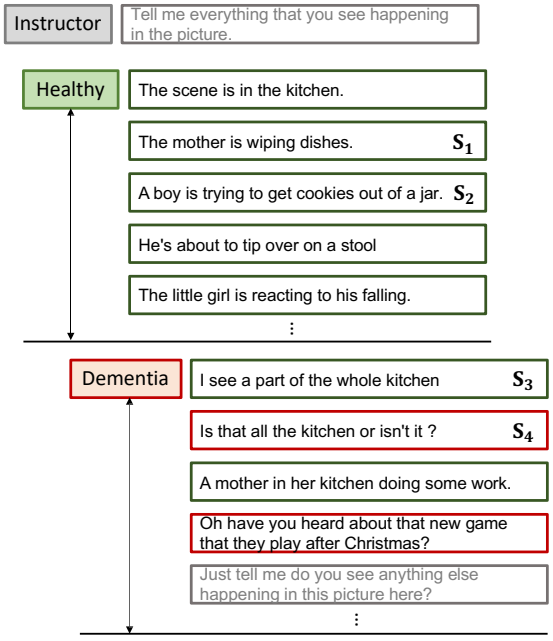

Figure 1: Snapshots from healthy controls and people with dementia describing the Cookie Theft Picture. Green frames indicate logically consistent utterances and red disruptive ones (e.g., elaborations or 'flight of ideas').

Fig. 1 illustrates two snapshots from people with dementia and healthy controls in the Pitt Corpus (Becker et al., 1994), containing subjects' descriptions of the Cookie Theft Picture (CTP, Appx. A) from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001). As shown in Fig. 1, dementia subjects present more disruptions in the logical consistency of their CTP narratives than healthy controls. For example, the pair of semantically unrelated utterances $\{S_1, S_2\}$ is logically consistent and descriptive. By contrast, even though $\{S_3, S_4\}$ are semantically related, the pair is logically inconsistent since the latter utterance disrupts the description of the CTP. Here we focus on learning coherence as logical-thematic consistency of utterances in narratives, rather than the semantic relatedness of entities across sentences, to capture

disruptive utterances, such as *flight of ideas* and *discourse elaborations*. The latter have been shown to be indicative of cognitive disorders (Abdalla et al., 2018; Iter et al., 2018).

The use of computational linguistics and natural language processing (NLP) to screen and monitor dementia progression has become an emergent and promising field (Fraser et al., 2016; König et al., 2018). However, recent work used language to distinguish people with Alzheimer's Disease (AD) from healthy controls, neglecting the longitudinal and fine-grained aspects of subjects' language impairments (Luz et al., 2020, 2021; Nasreen et al., 2021a). Here, we address this limitation by first learning the logical-thematic coherence of adjacent utterances in narratives, and then investigating the connection between longitudinal changes in language coherence and cognitive status.

Recent work for coherence in text has exploited deep (Cui et al., 2017; Feng and Mostow, 2021), discriminative (Xu et al., 2019), and generative (Laban et al., 2021) neural models for three evaluation tasks namely: a) the shuffle task (i.e., to discriminate genuine from randomly shuffled text), b) sentence ordering (i.e., to produce the correct order of sentences in a text) , and c) insertion (i.e., to predict the position of a missing sentence in a text). However these tasks are prone to learning the shuffle-ness of a text rather than its actual coherence (Laban et al., 2021). By contrast, our motivation is to learn the logical consistency of adjacent utterances in narratives to capture fine-grained coherence impairments (Fig. 1) rather than semantic relatedness or the global aspects of utterances' order. In this paper we make the following contributions:

- We define the new task of learning logical thematic coherence scores on the basis of the logical-thematic consistency of adjacent utterances (Sec. 3.1). We train on narratives from healthy controls in the DementiaBank Pitt Corpus (Becker et al., 1994), hypothesising that controls produce a logically consistent order of utterances. We investigate a range of state-of-the-art (SOTA) neural approaches and obtain models in three different settings: a) fine-tuning transformer-based models, b) fully training discriminative models, and c) zero-shot learning with transformer-based generative models (Sec. 3.3). Our experiments show that a fine-tuned transformer model (RoBERTa) achieves the highest discrimination between adjacent and non-adjacent utterances within a healthy cohort (Sec. 4.1.1).
- We introduce a human-interpretable digital coherence marker for dementia screening and monitoring from longitudinal language data. We first obtain logical thematic coherence scores of adjacent utterances and then aggregate these across the entire narrative (Sec. 3.1).
- We conduct a comprehensice longitudinal analysis to investigate how the digital coherence marker differs across healthy and dementia cohorts. The resulting digital coherence marker yields significant discrimination across healthy controls, people with mild cognitive impairment (MCI), and people with AD (Sec. 4.2.1).
- We compare our digital coherence marker against one based on semantic similarity, showing superior performance of the former in both distinguishing across cohorts (Sec. 4.2.1) and in detecting human-annotated disruptive utterances (Sec. 4.2.2).
- We evaluate our logical thematic coherence marker against three clinical bio-markers for cognitive impairment, showing high association and generalisability potential (Sec. 4.2.3).

2 Related Work

NLP and dementia: Early NLP work for dementia detection analysed aspects of language such as lexical, grammatical, and semantic features (Ahmed et al., 2013; Orimaye et al., 2017; Kavé and Dassa, 2018), and studied para-linguistic features (Gayraud et al., 2011; López-de Ipiña et al., 2013; Pistono et al., 2019). Recent work in this area has made use of manually engineered features (Luz et al., 2020, 2021; Nasreen et al., 2021a), disfluency features (Nasreen et al., 2021b; Rohanian et al., 2021), or acoustic embeddings (Yuan et al., 2020; Shor et al., 2020; Pan et al., 2021; Zhu et al., 2021). Closer to the current study, Abdalla et al. (2018) investigated discourse structure in people with AD by analyzing discourse relations. All such previous work has focused on differentiating across cohorts at fixed points in time without considering language changes over time.

Coherence modeling: The association between neuropsychological testing batteries and language leadresearchers to exploit linguistic features and naive approaches for capturing coherence in spontaneous speech to predict the presence of a

broad spectrum of cognitive and thought disorders. (Elvevåg et al., 2007; Bedi et al., 2015; Iter et al., 2018). Other work for coherence in text focused on feature engineering to implement some of the intuitions of Centering Theory (Lapata et al., 2005; Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Guinaudeau and Strube, 2013). Despite their success, existing models either capture semantic relatedness or entity transition patterns across sentences rather than logical-thematic consistency. **Neural coherence:** Driven by the success of deep neural networks, researchers exploited distributed sentences Cui et al. (2017), discriminative Xu et al. (2019), and BERT-based Feng and Mostow (2021) models by evaluating coherence mostly on the shuffle task (refer to Sec. 1 for more details). Recent work has shown that a zero-shot setting in generative transformers can be more effective than fine-tuning BERT or RoBERTa achieving a new SOTA performance for document coherence (Laban et al., 2021). Here, we investigate a variety of such successful architectures to learn the temporal logical-thematic consistency of utterances in transcribed narratives.

3 Methodology

3.1 Logical Thematic Coherence

Let us denote a collection C of N transcribed narratives from healthy controls, i.e., $C = \{d_k\}_{k=1}^{N}$, where each narrative consists of a sequence of utterances $\{u_i\}$. The logical thematic coherence task consists in learning scores from adjacent pairs of utterances (u_i, u_{i+1}) in the healthy controls, so that these are higher than corresponding non-adjacent pairs of utterances (u_i, u_j) in a narrative, where u_j is any forward utterance following the adjacent pair (Feng and Mostow, 2021)

To monitor changes in cognition over time, we define a digital language coherence marker by computing the logical thematic coherence scores of adjacent utterances in people with dementia and controls in a test set and aggregating these over the entire narrative. To obtain comparisons across cohorts, we calculate longitudinal changes in the coherence marker from the last to the first and between adjacent subjects' narratives over the study. To assess the reliability of the coherence marker, we compute changes in the coherence marker and in widely used clinical markers from the end to the beginning of the study.

3.2 Data

We have conducted experiments and trained coherence models on the DementiaBank Pitt Corpus (Becker et al., 1994), where subjects are asked to describe the Cookie Theft picture (Goodglass et al., 2001) up to 5 times across a longitudinal study (see Appx. B for more details about the Pitt Corpus). **Coherent pairs:** We have learnt the temporal logical-thematic coherence of adjacent utterances from the healthy cohort, consisting of 99 people with a total amount of 243 narratives. **Incoherent pairs:** We use logically inconsistent utterance ordering by choosing utterances following an adjacent pair, from the same narrative so as to avoid learning cues unrelated to coherence due to potential differences in language style (Patil et al., 2020; Feng and Mostow, 2021). While the level of coherence of controls may vary, we hypothesise that adjacent sentences by healthy controls will be more coherent than the negative instances, i.e. non-adjacent pairs from the same narrative. Table 1 summarizes the overall amount of utterances after splitting the healthy population into 80%, 10%, and 10% for training, validation, and testing.

Utterances	Training	Validation	Testing
# Coherent	2,178	223	233
# Incoherent	16,181	1,401	1,417

Table 1: Amount of coherent and incoherent utterances for learning logical thematic coherence from the healthy cohort.

To evaluate the ability of the digital language coherence marker to differentiate across cohorts and its reliability against the clinical bio-markers, we filtered people with dementia who have at least two narratives across the longitudinal study. This resulted in 62 people with AD and 14 people with MCI, with a total of 148 and 42 narratives respectively. We also included healthy controls, a total of 19 people with a total of 25 narratives.

3.3 Coherence Models

Baseline Digital Marker: We use Incoherence Model (Iter et al., 2018), which scores adjacent pairs of utterances in a narrative based on the cosine similarities of their sentence embeddings (Reimers and Gurevych, 2019). We consider three main neural architectures, known to achieve SOTA performance on document coherence, to learn logical thematic coherence: A) fine-tuning transformer-based

models, B) fully training discriminative models, and C) zero-shot learning with generative models.

Transformer-based Models: We fine-tune pre-trained transformers by maximising the probability that the second utterance in a pair follows the first (see Fig. 3 (A) in Appx. C). The model's input is a sequence of tokens in the form of $[CLS] + Utterance_1 + [SEP] + Utterance_2$, where $(Utterance_1, Utterance_2)$ is a pair of either coherent of incoherent utterances in a narrative (see Sec. 3.2), $[SEP]$ is an utterance separator token, and $[CLS]$ is a pair-level token, used for computing the coherence score. We append to the transformer module a feed-forward neural network (FFNN) followed by a sigmoid function where the coherence score f is the sigmoid function of FNNN that scales the output between 0 and 1. We fine-tune the models with a standard binary cross-entropy loss function (i.e., BCELoss), setting the output of the model to 1 for coherent and 0 for incoherent pairs of utterances.

We have experimented with the following variants: a) BERT-base (Lee and Toutanova, 2018) since it has been pre-trained on the Next Sentence Prediction (NSP) task which is similar to the task of scoring the coherence of adjacent utterances. b) RoBERTa-base (Liu et al., 2019), which has been pre-trained without the NSP task. c) a Convolutional Neural Network baseline (Cui et al., 2017) which uses pre-trained word embeddings extracted by BERT-base (refer to Appx. C for a detailed description).

Discriminative Models: We have trained discriminative models by maximizing the probability of an utterance pair being coherent. We have experimented with an architecture previously shown effective in coherence modelling for both speech (Patil et al., 2020) and text. (Xu et al., 2019).

The model receives a pair of utterances and a sentence encoder maps the utterances to real-value vectors U_1 and U_2 (see Fig. 3 (B) in Appx. C). The model then computes the concatenation of the two encoded utterances, as follows:

$$concat[U_1, U_2, U_1 - U_2, U_1 * U_2, |U_1 - U_2|] \quad (1)$$

, where $U_1 - U_2$ is the element-wise difference, $U_1 * U_2$ is the element-wise product, and $|U_1 - U_2|$ is the absolute value of the element-wise difference between the two encoded utterances. The choice to represent the difference between utterances in the form of Eq. 1 was introduced by Xu et al. (2019) as a high level statistical function that could capture local level interaction between utterances and we make the same assumption. Finally, the concatenated feature representation is fed to a one-layer MLP to output the coherence score f. We have trained the model in bi-directional mode with inputs (U_1, U_2) and (U_2, U_1) for the forward and backward operations and used a margin loss as follows:

$$L(f^+, f^-) = max(0, n - f^+ + f^-) \quad (2)$$

, where f^+ is the coherence score of a coherent pair of utterances, f^- thescore of an incoherent pair, and n the margin hyperparameter. The model can work with any pre-trained sentence encoder. Here, we experiment with two variants: a) pre-trained sentence embeddings from SentenceBERT (Reimers and Gurevych, 2019)(**DCM-sent**), and b) averaged pre-trained word embeddings extracted from BERT-base (Lee and Toutanova, 2018)(**DCM-word**).

Generative Models: We experiment with a zero-shot setting for generative transformers, an approach that previously achieved best out-of-the-box performance for document coherence (Laban et al., 2021). We provide a pair of utterances to a generative transformer and compute the perplexity in the sequence of words for each pair (refer to Appx. C for a detailed description). Perplexity is defined as the exponential average log-likelihood in a sequence of words within a pair P as follows:

$$PPL(P) = exp\Big\{-\frac{1}{t}\sum_{i}^{t} p(w_i|w_{<i})\Big\}, \quad (3)$$

, where $p(w_i|w_{<i})$ is the likelihood of the i^{th} word given the preceding words $w_{<i}$ within a pair of utterances. Finally, we approximate the coherence score f as follows:

$$f = 1 - PPL(P), \quad (4)$$

We use $1 - PPL$ rather than PPL since low perplexity indicates that a pair is likely to occur, but we need high coherence scores for sequential pairs.

We have experimented with two SOTA generative transformers, of different sizes and architecture: a) **GTP2**, a decoder transformer-based model (Radford et al., 2019) and b) **T5**, an encoder-decoder transformer-based model (Raffel et al.,

2020). In the end we also pre-train T5-base, i.e., **T5-base**$_{pre}$. In particular, we feed sequential pairs of utterances and consider the loss on the second sequential sentence within the pair, just like sequence to sequence models. For testing, we extract coherence scores according to Eq. 4 for coherent and incoherent pairs.

For the training details of coherence models please refer to Appx. F.

3.4 Evaluation Metrics

For evaluating the temporal logical thematic coherence models, we report the average coherence score of adjacent and non-adjacent utterance pairs, denoted as f^+ and f^-, respectively. The higher the f score, the more coherent the pair. We also report the models' accuracy on adjacent utterances denoted as *temporal* accuracy, i.e., Acc_{temp}, calculated as the correct rate between the adjacent utterances recognized as coherent and the total number of adjacent pairs in the test corpus. In particular, a pair of adjacent utterances $\{u_i, u_{i+1}\}$ in the test set is perceived as coherent if its coherence score $f_{(u_i,u_{i+1})}$ is higher than the coherence score $f_{(u_i,u_{k>i+1})}$ of the corresponding non-adjacent pair of utterances as follows:

$$f(u_i, u_{i+1}) = \begin{cases} 1 & \text{if } f_{(u_i,u_{i+1})} > f_{(u_i,u_{k>i+1})} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

, where 1 corresponds to coherent and 0 to incoherent pair, correspondingly. The coherence across an entire narrative is approximated by averaging the coherence scores of adjacent utterances, denoted as *entire* accuracy, i.e., Acc_{entire}. Similarly, the entire accuracy is calculated as the correct rate of narratives recognized as coherent out of the total amount of narratives in the test corpus. A narrative is perceived as coherent if the averaged scores of the adjacent utterances are higher than the average scores of the non-adjacent ones within a narrative. The higher the temporal and entire accuracy, the better the model. Finally, we report the absolute percentage difference in f scores between adjacent and non-adjacent utterances, denoted $\%\Delta$ (refer to Appx. D for more details), and the averaged loss of the models. The higher and more significant the $\%\Delta$, the better the model, while the reverse holds for the averaged loss.

To investigate the reliability of the digital coherence marker, we evaluate against three different clinical bio-markers collected from people with dementia. These are the Mini-Mental State Examination (MMSE), the Clinical Dementia Rating (CDR) scale (Morris, 1997), and the Hamilton Depression Rating (HDR) scale (Williams, 1988). The lower the MMSE score the more severe the cognitive impairment. The opposite is true of the other scores, where a higher CDR score denotes more severe cognitive impairment and higher HDR scores indicate more severe depression (for more details about the bio-markers please refer to Appx. E).

4 Experimental Results

4.1 Logical Thematic Coherence Models

4.1.1 Quantitative Analysis

Table 2 summarizes the performance of logical thematic coherence models trained on the healthy cohort. Overall, fine-tuned transformerssignificantly outperform discriminative and generative transformer models. All models score higher on consecutive utterance pairs than non-consecutive ones. While the absolute percentage difference of coherence scores between sequential and non-sequential pairs of utterances is higher for the discriminative models, $\%\Delta$ has a higher significance for the transformer-based models.

BERT and RoBERTa are the best performing models, achieving a significant high entire accuracy (100%), meaning that the model is able to predict all the narratives in the healthy population as being coherent, in line with our hypothesis. RoBERTa yielded an increased logical thematic coherence accuracy of 81.4% compared to 75.4% for BERT. Despite the original BERT being trained with two objectives, one of which is Next Sentence Prediction (NSP), an indirect signal for the coherence of adjacent utterances, RoBERTa, trained without the NSP objective, outperformed BERT. Presumably, RoBERTa outperforms BERT since the former was trained on a much larger dataset and using a more effective training procedure. Moreover, the simple CNN baseline, while performing worse than BERT and RoBERTa still outperforms the discriminative and generative models, which shows the effectiveness of fine-tuning.

The discriminative models perform better when using pre-trained embeddings from BERT rather than pre-trained sentence embeddings. Our experiments show that discriminative models are outperformed by transformers when modelling thematic logical coherence in transcribed narratives. This

Model	Setting	Avg. f^+	Avg. f^-	$\%\Delta$	Avg. Acc_{temp}	Avg. Acc_{entire}	Avg. Loss
CNN	Training	0.560	0.475	18.2†	73.4%	92.0%	0.636
BERT-base	Fine-tuning	0.630	0.422	49.1†	75.4%	**100.0%**	0.575
RoBERTa-base	Fine-tuning	0.604	0.353	**71.0**†	**81.4%**	**100.0%**	**0.554**
DCM-sent	Training	-0.034	-1.975	**98.2**†	63.9%	76.0%	3.64
DCM-word	Training	0.282	-1.068	**126.4**†	69.6%	80.0%	3.84
GPT2-base	Zero Shot	-383.8	-384.8	0.3	50.4%	48.0%	-
GPT2-medium	Zero Shot	-313.0	-318.5	1.7	48.9%	48.0%	-
GPT2-large	Zero Shot	-290.1	-298.8	-2.9	50.0%	60.0%	-
T5-base	Zero Shot	-0.668	-0.751	11.0	64.8%	64.0%	-
T5-large	Zero Shot	-3.674	-3.996	8.1	58.2%	60.0%	-
T5-base$_{pre}$	Pre-train	-0.224	-0.208	7.3	46.1%	40.0%	0.376

Table 2: Performance of logical thematic coherence models trained on healthy controls in three different settings; A) training, B) fine-tuning, and C) zero-shot. f^+ is the coherence score of adjacent utterances, f^- the coherence score of non-adjacent ones, and $\%\Delta$ the absolute percentage difference between f^+ and f^-. $\dagger$ denotes significant difference between the two coherence scores. Acc_{temp} and Acc_{entire} measure accuracy on adjacent utterances and entire narratives, respectively. Best performance is highlighted in bold.

is contrary to earlier work (Xu et al., 2019; Patil et al., 2020) where discriminative models outperformed early RNN based models, but we note that this work did not compare against transformers.

Despite Laban et al. (2021) showing that a zero-shot setting in generative transformers can be more effective than fine-tuning BERT or RoBERTa, our experiments show that this setting has the worst performance. The results did not improve even when we pre-trained the T5 model on the Pitt corpus (see T5-base$_{pre}$ in Table 2). We presume that large pre-trained language models may suffer from domain adaptation issues here and operate on too short a window to capture logical consistency in narratives. Future work could investigate fine-tuning or prompt-training generative transformers for this task.

4.2 The digital Language Coherence Marker

Here, we exploited the best-performing logical thematic coherence model, i.e., RoBERTa, to obtain a digital language coherence marker for subjects across different cohorts over the longitudinal study (refer to Sec. 3.1 for more details). We first present results regarding the longitudinal discrimination ability for this marker and then show its reliability by evaluating against three clinical bio-markers.

4.2.1 Longitudinal Discrimination Ability

We analyzed changes in the digital marker over time and across cohorts. First, we calculated the average of digital markers across the three cohorts. The column $Marker$ in Table 3 summarizes the results. The averaged digital marker was higher in the healthy cohort than in MCI and AD cohorts. Similarly, the averaged marker in the MCI group was higher than that in the AD group. However, the difference was significant only between the healthy and AD cohorts ($p < 0.05$) [1].

We subsequently calculated changes in the digital marker from the end to the start of the study and across the cohorts (i.e., $\Delta_{(end-onset)}$ in Table 3). There was a significant decrease for the MCI and AD groups and a significant increase for the healthy controls ($p < 0.05$) [1]. The increase in healthy controls is presumably because subjects are able to remember and do better at the CTP description when seeing it again (Goldberg et al., 2015). Moreover, we noticed that people with MCI exhibited more substantial change than those with AD, despite the average digital coherence marker of the former being 0.597 compared to 0.567 for the latter.

We also calculated changes in the digital marker between adjacent narratives over time and then aggregated the changes within subjects in the study. In Table 3, we report the average change across cohorts, i.e., $\Delta_{(long)}$. We obtain similar results as the ones taken from end to start.

We finally compared the longitudinal discrimination ability of our proposed digital marker with a baseline digital marker based on the semantic relatedness of adjacent utterances (refer to Sec. 3.3).

[1] We use a nonparametric test, namely the Mann-Whitney test, to measure if the distribution of a variable is different in two groups.

	Our digital marker			Baseline digital marker		
Cohort	**Marker**	$\Delta_{(end-start)}$	$\Delta_{(long)}$	**Marker**	$\Delta_{(end-start)}$	$\Delta_{(long)}$
Healthy	**0.604 (0.08)**	**0.09 (0.07)**	**0.07 (0.05)**	0.249 (0.05)	0.02 (0.06)	0.01(0.06)
MCI	0.597 (0.09)	**-0.05 (0.09)**	**-0.05 (0.07)**	0.262 (0.06)	-0.03 (0.07)	-0.03 (0.06)
AD	**0.567 (0.10)**	**-0.02 (0.16)**	**-0.02 (0.11)**	0.241 (0.07)	-0.01 (0.08)	-0.01 (0.06)

Table 3: Longitudinal discrimination ability between the proposed digital marker and a baseline based on semantic similarity. Marker: Average of coherence marker within a population. $\Delta_{(end-start)}$: Average change of the marker from the end to the beginning of the study. $\Delta_{(long)}$: Average change of the digital marker between adjacent narratives within subjects. Numbers in () refer to corresponding standard deviations. Numbers in bold denote significant difference between the health controls and dementia cohorts (see Sec. 4.2.1).

The averaged baseline marker was higher in the MCI cohort than in healthy and AD cohorts (see Table 3). Moreover, there was no significant difference across the cohorts. On the other hand, we observed similar changes (i.e., $\Delta_{(end-start)}$ and $\Delta_{(long)}$ in Table 3) in the baseline marker over time compared to the one proposed in this paper. However, such changes were not significant across cohorts for the baseline marker ($p > 0.05$) [1].

4.2.2 Evaluation on Human-Annotated Disruptive Utterances

We investigated the effectiveness of the digital coherence marker in capturing disruptive utterances in narratives, and compared it with the baseline digital marker. Such disruptive utterances are annotated with the code $[+\ exc]$ in the transcripts of the Pitt corpus and constitute a significant indicator of AD speech (Abdalla et al., 2018; Voleti et al., 2019). Out of 1,621 pairs of adjacent utterances in the AD cohort, 543 ones (33%) are disruptive. For the baseline marker, the average score of disruptive utterances decreased to 0.19 (STD=0.17) compared to 0.26 (STD=0.17) for non-disruptive ones, i.e., an absolute percentage difference [2] of 31%. For our proposed marker, the average score of disruptive utterances decreased to 0.41 (STD=0.09) from 0.64 (STD=0.15) for non-disruptive ones, i.e., an absolute percentage difference of 44%. The results showed that both digital markers significantly captured disruptive utterances ($p_{t-test} < 0.05$). However, our proposed digital marker is more robust in capturing such utterances.

4.2.3 Association with Clinical Bio-markers

We investigated the reliability of the digital marker by associating its changes with different degrees of changes in cognitive status from the end to the beginning of the longitudinal study, as expressed by widely accepted cognition scales. We analyzed association patterns in the largest cohort, i.e., the AD group consisting of 62 participants.

We first investigated the association between changes in the coherence marker against the Mini-Mental State Examination (MMSE) (Morris, 1997). MMSE ranges from 0-30. The higher the MMSE score, the higher the cognitive function (refer to Appx. E for more details about MMSE). Here, we have split the AD population into four bins on the basis of the magnitude of MMSE change. Table 4 provides details regarding bin intervals and the association of changes between the MMSE and the digital coherence marker.

Bin	**# Subjects**	**Δ MMSE**	**Δ Coherence**
Low	25	[-6,2]	-0.003 (0.089)
Minor	17	[-12,-7]	-0.030 (0.094)
Moderate	11	[-18,-13]	-0.076 (0.095)
Severe	9	[-27,-19]	-0.200 (0.104)

Table 4: Association between changes in Mini-Mental State Examination (MMSE) and the digital coherence marker in AD patients at different degrees of cognitive decline. Numbers in [,] define the lower and upper values of each bin interval. Numbers in () refer to the standard deviation. # Subjects = Population within bins. Δ = Change from the end to the onset of the study.

Overall, we observe that the digital marker decreases across the population for the different degrees of cognitive decline. In particular, the higher the difference in MMSE, the more substantial the decrease in the digital marker change over the longitudinal study. For people with moderate or severe cognitive decline, the coherence decreased significantly compared to that of people with low cognitive decline ($p < 0.05$) [1,3].

[2] For the definition refer to 3.4.

[3] Here, we investigated how coherence change distribu-

Next, we investigated the association between changes in the coherence marker and the Clinical Dementia Rating (CDR) (Morris, 1997). CDR is based on a scale of 0–3 in assessing people with dementia. The higher the CDR, the lower the cognitive function (refer to Appx. E for more details about CDR). Here, we split the AD population into low, minor, moderate and severe bins according to the magnitude of CDR change, i.e., Δ CDR in Table 5. The higher the CDR change the more severe the cognitive decline over time.

Bin	# Subjects	Δ CDR	Δ Coherence
Low	20	[0, 0.5]	-0.009 (0.091)
Minor	16	(0.5,1.5]	-0.011 (0.060)
Moderate	15	(1.5,2.5]	-0.060 (0.110)
Severe	11	(2.5,3]	-0.125 (0.078)

Table 5: Association between changes in Clinical Dementia Rating (CDR) and the digital coherence marker in AD patients atdifferent degrees of cognitive decline. Numbers in $(,]$ define the lower and upper values of each bin interval. Numbers in $()$ refer to the standard deviation. # Subjects = Population within bins. Δ = Change from the end to the onset of the study.

The digital coherence marker decreased across the population at different degrees of CDR change. In particular, the higher the increase in CDR, the higher the decrease in the digital coherence marker over the longitudinal study. Changes in the digital coherence marker are similar for people with low and minor cognitive decline. However, there is significant decrease in coherence for the moderate and severe bins compared to the minor and mild ones $p < 0.05$) [1,3] .

Finally, we investigated the association between the coherence marker with the Hamilton Depression Rating (HDR) (Williams, 1988). HDR is one of the most widely used and accepted instruments for assessing depression. It is based on a 17-item scale. The higher the HDR, the more severe the level of depression (refer to Appx. E for more details about HDR). We investigated associations between the last HDR record [4] and changes in the digital coherence marker from the end to start of the study. Table 6 summarizes the association between HDR and changes in the digital coherence marker. Changes in coherence were similar for people with no or mild depression. However, there was a significant decrease for people with moderate depression ($p < 0.05$) [1,3]. This is in line with current studies showing that individuals experiencing difficulty constructing coherent narratives generally report low well-being and more depressive symptoms (Vanderveren et al., 2020).

Bin	# Subjects	HDR	Δ Coherence
No Depression	17	[0,7]	-0.02 (0.11)
Mild	18	[8,16]	-0.01 (0.10)
Moderate	14	[17,23]	-0.21 (0.10)

Table 6: Association between the last Hamilton Depression Rating (HDR) record and changes in the digital coherence for AD patients. Numbers in $[,]$ define the lower and upper values of each bin interval. Numbers in $()$ refer to the standard deviation. # Subjects = Population within bins. Δ = Change from the end to the onset of the study.

5 Conclusion

We have introduced a new task for modelling the logical-thematic temporal coherence of utterances in short transcribed narratives to capture disruptive turns indicative of cognitive disorders. To this end, we have investigated transformer-based, discriminative, and generative neural approaches. Our experiments show that a fine-tuned transformer model (RoBERTa) achieves the best performance in capturing the coherence of adjacent utterances in narratives from the healthy cohort. We aggregate temporal language coherence to create a human-interpretable digital language coherence marker for longitudinal monitoring of cognitive decline. Longitudinal analysis showed that the digital marker is able to distinguish people with mild cognitive impairment, those with Alzheimer's Disease (AD) and healthy controls. A comparison with a baseline digital marker based on semantic similarity showed the superiority of our digital marker. Moreover, evaluation against three clinical bio-markers showed that language coherence can capture changes at different degrees of cognitive decline and achieves significant discrimination between people with moderate or severe cognitive decline within an AD population. It can also capture levels of depression, showing generalisability potential. In future, we aim to integrate disfluency language patterns and develop strategies for improving the performance of generative models.

tions differ across the AD population at different degrees of cognitive decline progression.

[4]We considered the last HDR record instead of changes in HDR over time since there were missing HDR measurements in the study.

Limitations

Monitoring dementia using computational linguistics approaches is an important topic. Previous work has mostly focused on distinguishing people with AD from healthy controls rather than monitoring changes in cognitive status per individual over time. In this study, we have used the Pitt corpus, currently the largest available longitudinal dementia dataset, to investigate longitudinal changes in logical coherence and their association with participants' cognitive decline over time. An important limitation of the Pitt corpus is that the longitudinal aspect is limited, spanning up to 5 sessions/narratives maximum per individual with most participants contributing up to two narratives. Moreover, the number of participants is relatively small, especially for the MCI cohort. In the future, we aim to address these limitations by investigating the generalisability of the proposed digital language coherence marker on a recently introduced rich longitudinal dataset for dementia (currently under review) and on transcribed psychotherapy sessions (data is collected in Hebrew) to monitor mood disorders.

In this study, we used manually transcribed data from Pitt. In a real-world scenario, participants mostly provide speech via a speech elicitation task. This implies that the introduced method requires an automatic speech recognition (ASR) system robust to various sources of noise to be operationalized. ASR for mental health is currently underexplored, with most transcription work being done by human transcription.

It may be that the proposed digital coherence marker becomes a less accurate means for monitoring dementia when people experience other comorbidities, neurodegenerative and mental illnesses, that significantly affect speech and language. Indeed, cognitive-linguistic function is a strong biomarker for neuropsychological health (Voleti et al., 2019).

Finally, there is a great deal of variability to be expected in speech and language data affecting the sensitivity of the proposed digital marker. Both speech and language are impacted by speaker identity, context, background noise, spoken language etc. Moreover, people may vary in their use of language due to various social contexts and conditions, a.k.a., style-shifting (Coupland, 2007). Both inter and intra-speaker variability in language could affect the sensitivity of the proposed digital marker. While it is possible to tackle intra-speaker language variability, e.g., by integrating speaker-dependent information to the language, the inter-speaker variability remains an open-challanged research question.

Ethics Statemen

Our work does not involve ethical considerations around the analysis of the DementiaBank Pitt corpus as it is widely used. Ethics was obtained by the original research team by James Backer and participating individuals consented to share their data in accordance with a larger protocol administered by the Alzheimer and Related Dementias Study at the University of Pittsburgh School of Medicine (Becker et al., 1994). Access to the data is password protected and restricted to those signing an agreement.

This work uses transcribed dementia data to identify changes in cognitive status considering individuals' language . Potential risks from the application of our work in being able to identify cognitive decline in individuals are akin to those who misuse personal information for their own profit without considering the impact and the social consequences in the broader community. Potential mitigation strategies include running the software on authorised servers, with encrypted data during transfer, and anonymization of data prior to analysis. Another possibility would be to perform on device processing (e.g. on indidivuals' computers or other devices) for identifying changes in cognition and that the results of the analysis would only be shared with authorised individuals. Individuals would be consented before any of our software would be run on their data.

References

Mohamed Abdalla, Frank Rudzicz, and Graeme Hirst. 2018. Rhetorical structure and alzheimer's disease. *Aphasiology*, 32(1):41–60.

Samrah Ahmed, Anne-Marie F Haigh, Celeste A de Jager, and Peter Garrard. 2013. Connected speech as a marker of disease progression in autopsy-proven alzheimer's disease. *Brain*, 136(12):3727–3737.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

James T Becker, François Boiler, Oscar L Lopez, Judith Saxton, and Karen L McGonigle. 1994. The natural history of alzheimer's disease: description of

study cohort and accuracy of diagnosis. *Archives of neurology*, 51(6):585–594.

Gillinder Bedi, Facundo Carrillo, Guillermo A Cecchi, Diego Fernández Slezak, Mariano Sigman, Natália B Mota, Sidarta Ribeiro, Daniel C Javitt, Mauro Copelli, and Cheryl M Corcoran. 2015. Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia*, 1(1):1–7.

Nikolas Coupland. 2007. *Style: Language variation and identity*. Cambridge University Press.

Baiyun Cui, Yingming Li, Yaqing Zhang, and Zhongfei Zhang. 2017. Text coherence analysis based on deep neural network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2027–2030.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129.

Brita Elvevåg, Peter W Foltz, Daniel R Weinberger, and Terry E Goldberg. 2007. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316.

Jingrong Feng and Jack Mostow. 2021. Towards difficulty controllable selection of next-sentence prediction questions. In *EDM*.

Katrina E Forbes-McKay and Annalena Venneri. 2005. Detecting subtle spontaneous language decline in early alzheimer's disease with a picture description task. *Neurological sciences*, 26(4):243–254.

Kathleen C. Fraser, Jed A. Meltzer, and Frank Rudzicz. 2016. Linguistic features identify Alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49(2):407–422.

Frederique Gayraud, Hye-Ran Lee, and Melissa Barkat-Defradas. 2011. Syntactic and lexical context of pauses and hesitations in the discourse of alzheimer patients and healthy elderly subjects. *Clinical linguistics & phonetics*, 25(3):198–209.

Terry E Goldberg, Philip D Harvey, Keith A Wesnes, Peter J Snyder, and Lon S Schneider. 2015. Practice effects due to serial cognitive assessment: implications for preclinical alzheimer's disease randomized controlled trials. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, 1(1):103–111.

H. Goodglass, E. Kaplan, S. Weintraub, and B. Barresi. 2001. The boston diagnostic aphasia examination. *Philadelphia, PA: Lippincott, Williams & Wilkins*.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103.

Paul Hoffman, Lucy Cogdell-Brooke, and Hannah E Thompson. 2020. Going off the rails: Impaired coherence in the speech of patients with semantic control deficits. *Neuropsychologia*, 146:107516.

Dan Iter, Jong Yoon, and Dan Jurafsky. 2018. Automatic detection of incoherent speech for diagnosing schizophrenia. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 136–146.

Gitit Kavé and Ayelet Dassa. 2018. Severity of alzheimer's disease and language features in picture descriptions. *Aphasiology*, 32(1):27–40.

Alexandra König, Nicklas Linz, Johannes Tröger, Maria Wolters, Jan Alexandersson, and Philippe Robert. 2018. Fully automatic speech-based analysis of the semantic verbal fluency task. *Dementia and Geriatric Cognitive Disorders*, 45(3-4):198–209.

Philippe Laban, Luke Dai, Lucas Bandarkar, and Marti A Hearst. 2021. Can transformer models measure coherence in text? re-thinking the shuffle test. *arXiv preprint arXiv:2107.03448*.

Mirella Lapata, Regina Barzilay, et al. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, volume 5, pages 1085–1090. Citeseer.

J Devlin M Chang K Lee and K Toutanova. 2018. Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Karmele López-de Ipiña, Jesus-Bernardino Alonso, Carlos Manuel Travieso, Jordi Solé-Casals, Harkaitz Egiraun, Marcos Faundez-Zanuy, Aitzol Ezeiza, Nora Barroso, Miriam Ecay-Torres, Pablo Martinez-Lage, et al. 2013. On the selection of non-invasive methods based on speech analysis oriented to automatic alzheimer disease diagnosis. *Sensors*, 13(5):6730–6745.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's dementia recognition through spontaneous speech: the adress challenge. *arXiv preprint arXiv:2004.06833*.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting cognitive decline using speech only: The adresso challenge. *arXiv preprint arXiv:2104.09356*.

G McKhann. 1987. Diagnostics and statistical manual of mental disorders. *Arlington, VA: American Psychiatric Association*.

John C Morris. 1997. Clinical dementia rating: a reliable and valid diagnostic and staging measure for dementia of the alzheimer type. *International psychogeriatrics*, 9(S1):173–176.

Shamila Nasreen, Julian Hough, Matthew Purver, et al. 2021a. Detecting alzheimer's disease using interactional and acoustic features from spontaneous speech. Interspeech.

Shamila Nasreen, Morteza Rohanian, Julian Hough, and Matthew Purver. 2021b. Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. *Frontiers in Computer Science*, page 49.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. Predicting probable alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18(1):1–13.

Yilin Pan, Bahman Mirheidari, Jennifer M Harris, Jennifer C Thompson, Matthew Jones, Julie S Snowden, Daniel Blackburn, and Heidi Christensen. 2021. Using the outputs of different automatic speech recognition paradigms for acoustic-and bert-based alzheimer's dementia detection through spontaneous speech. In *Interspeech*, pages 3810–3814.

Rajaswa Patil, Yaman Kumar Singla, Rajiv Ratn Shah, Mika Hama, and Roger Zimmermann. 2020. Towards modelling coherence in spoken discourse. *arXiv preprint arXiv:2101.00056*.

Aurélie Pistono, Jeremie Pariente, C Bézy, B Lemesle, J Le Men, and Mélanie Jucla. 2019. What happens when nothing happens? an investigation of pauses as a compensatory mechanism in early alzheimer's disease. *Neuropsychologia*, 124:133–143.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Gisela Redeker. 2000. Coherence and structure in text and discourse. *Abduction, belief and context in dialogue*, 233(263).

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Morteza Rohanian, Julian Hough, and Matthew Purver. 2021. Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs. *arXiv preprint arXiv:2106.15684*.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*.

Elien Vanderveren, Loes Aerts, Sofie Rousseaux, Patricia Bijttebier, and Dirk Hermans. 2020. The influence of an induced negative emotional state on autobiographical memory coherence. *Plos one*, 15(5):e0232495.

Rohit Voleti, Julie M Liss, and Visar Berisha. 2019. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE journal of selected topics in signal processing*, 14(2):282–298.

Su Wang, Greg Durrett, and Katrin Erk. 2020. Narrative interpolation for generating and understanding stories. *arXiv preprint arXiv:2008.07466*.

Janet BW Williams. 1988. A structured interview guide for the hamilton depression rating scale. *Archives of general psychiatry*, 45(8):742–747.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019. A cross-domain transferable neural coherence model. *arXiv preprint arXiv:1905.11912*.

Jiahong Yuan, Yuchen Bian, Xingyu Cai, Jiaji Huang, Zheng Ye, and Kenneth Church. 2020. Disfluencies and fine-tuning pre-trained language models for detection of alzheimer's disease. In *INTERSPEECH*, volume 2020, pages 2162–6.

Youxiang Zhu, Abdelrahman Obyat, Xiaohui Liang, John A Batsis, and Robert M Roth. 2021. Wavbert: Exploiting semantic and non-semantic speech using wav2vec and bert for dementia detection. In *Interspeech*, pages 3790–3794.

A The Cookie Theft Picture

For the PD task, the examiner asks subjects to describe the picture (see Fig. 2) by saying, "Tell me everything you see going on in this picture". Then subjects might say, "there is a mother who is drying dishes next to the sink in the kitchen. She is not paying attention and has left the tap on. As a result, water is overflowing from the sink. Meanwhile, two children are attempting to make cookies from a jar when their mother is not looking. One of the children, a boy, has climbed onto a stool to get up to the cupboard where the cookie jar is stored. The

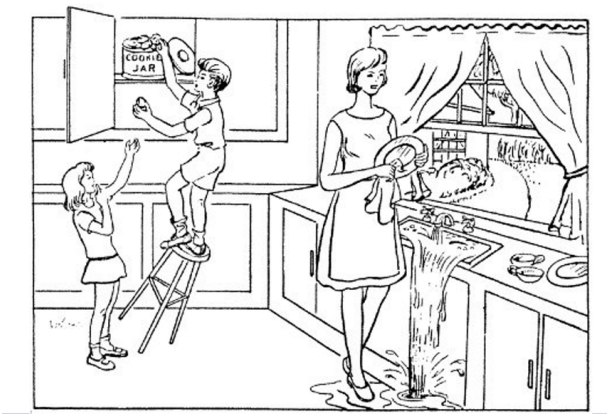

Figure 2: The Cookie Theft Picture from the Boston Diagnostic Aphasia Examination."

stool is rocking precariously. The other child, a girl, is standing next to the stool and has her hand outstretched ready to be given cookies.

B DementiaBank Pitt Corpus

The dataset was gathered longitudinally between 1983 and 1988 as part of the Alzheimer Research Program at the University of Pittsburgh. The study initially enrolled 319 participants according to the following eligibility criteria: all the participants were required to be above 44 years old, have at least seven years of education, have no history of major nervous system disorders, and have an initial Mini-Mental State Examination score above 10. Finally, the cohort consisted of 282 subjects. In particular, the cohort included 101 healthy control subjects (HC) and 181 Alzheimer's disease subjects (AD). An extensive neuropsychological assessment was conducted on the participants, including verbal tasks and the Mini-Mental State Examination (MMSE).

C Architecture Overview of Models

We consider three main types of coherence models, in three different settings: a) fine-tuning transformer-based models, b) fully training discriminative models, and c) zero-shot learning with transformer-based generative models . Fig. 3 provides the overall architecture of coherence models in each setting. The models receive a pair of utterances in the input and output a coherence score for the given pair. The main difference between the three is that discriminative models learn constrastive patterns to obtain the probability of an utterance pair being coherent while the transformer-based models maximise the probability of the second utterance in the pair following the first.

When we experiment with zero-shot learning (Fig. 3 (C)), we feed each generative transformer model with adjacent pair of utterances. For calculating the probability of each word given its preceding ones, i.e., context, we use cross-entropy loss, calculated between the genuine pair and the generated output. The exponentiation of the cross-entropy loss between the input and model predictions is equivalent to perplexity, defined as the exponentiated average negative log-likelihood of the tokenized sequence (see Eq. 3). A high perplexity implies a low model predictability. To this goal, we approximate the coherence as $1 - P$ (see Fig. 3 (C)).

For CNN (Cui et al., 2017), we use pre-trained word embeddings extracted by BERT. Each pair of utterances is transformed to a 2-dimensional matrix $\in \mathcal{R}^{d \times N}$, where d denotes the dimension of pre-trained BERT embeddings and N is the total number of words across the pair. The rest of the architecture is similar to that one we used for transformer-based models (see Fig. 3 (A)). In particular, we append to the CNN module a feed-forward neural network (FFNN) followed by a sigmoid function. The coherence score is the sigmoid function of FNNN that scales the output between 0 and 1. We trained the model by freezing the pre-trained BERT embeddings.

D Absolute Percentage Coherence Score Difference Formula

The absolute percentage difference in f scores equals the absolute value of the change in f between adjacent and non-adjacent sentences divided by the average of positive, i.e., f^+, and negative, i.e., f^-, coherence scores, all multiplied by 100, as follows:

$$\%\Delta f = \frac{|\Delta f|}{\left[\frac{\Sigma f}{2}\right]} \times 100 = \frac{|f^+ - f^-|}{\left[\frac{f^+ + f^-}{2}\right]} \times 100$$

The order of the coherence scores does not matter as we are simply dividing the difference between two scores by the average of the two coherence scores.

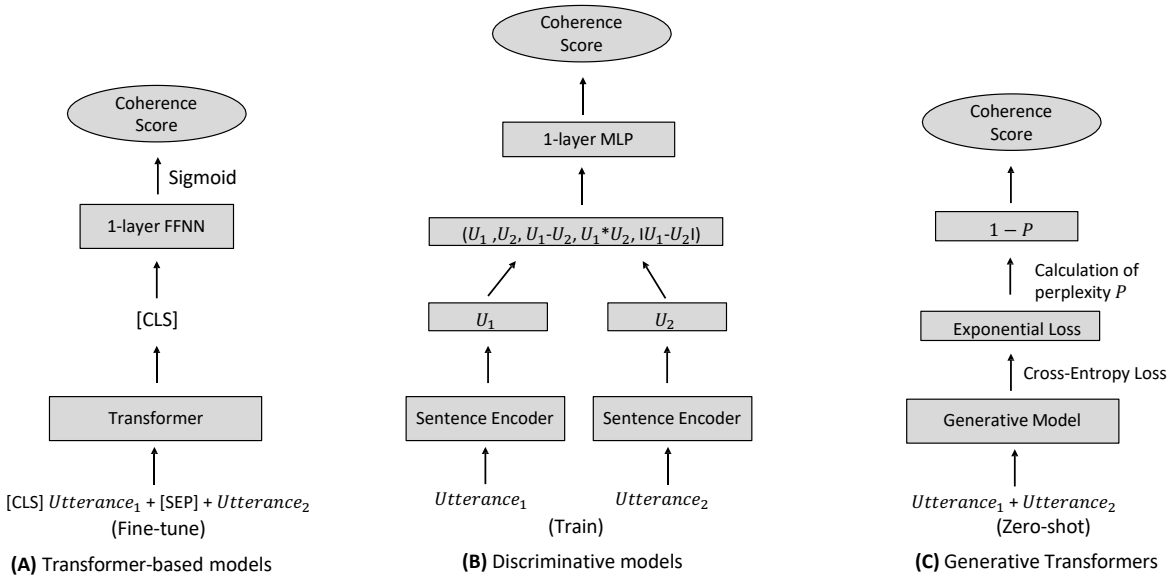

Figure 3: Architecture overview of coherence models in the three settings. The final output is always a coherence score for a given pair of sentences.

E Clinical Bio-Markers

E.1 Mini-Mental State Examination (MMSE)

The Mini-Mental State Examination (MMSE) has been the most common method for diagnosing AD and other neurodegenerative diseases affecting the brain. It was devised in 1975 by Folstein et al. as a simple standardized test for evaluating the cognitive performance of subjects, and where appropriate to qualify and quantify their deficit. It is now the standard bearer for the neuropsychological evaluation of dementia, mild cognitive impairment, and AD.

The MMSE was designed to give a practical clinical assessment of change in cognitive status in geriatric patients. It covers the person's orientation to time and place, recall ability, short-term memory, and arithmetic ability. It may be used as a screening test for cognitive loss or as a brief bedside cognitive assessment. By definition, it cannot be used to diagnose dementia, yet this has turned into its main purpose.

The MMSE includes 11 items, divided into 2 sections. The first requires verbal responses to orientation, memory, and attention questions. The second section requires reading and writing and covers ability to name, follow verbal and written commands, write a sentence, and copy a polygon. All questions are asked in a specific order and can be scored immediately by summing the points assigned to each successfully completed task; the maximum score is 30. A score of 25 or higher is classed as normal. If the score is below 24, the result is usually considered to be abnormal, indicating possible cognitive impairment. The MMSE has been found to be sensitive to the severity of dementia in patients with Alzheimer's disease (AD). The total score is useful in documenting cognitive change over time.

E.2 Clinical Dementia Rating (CDR)

The Clinical Dementia Rating (CDR) is a global rating device that was first introduced in a prospective study of patients with mild "senile dementia of AD type" (SDAT) in 1982 (Hughes et al., 1982). New and revised CDR scoring rules were later introduced (Berg, 1988; Morris, 1993; Morris et al., 1997). CDR is estimated on the basis of a semistructured interview of the subject and the caregiver (informant) and on the clinical judgment of the clinician. CDR is calculated on the basis of testing six different cognitive and behavioral domains such as memory, orientation, judgment and problem solving, community affairs, home and hobbies performance, and personal care. The CDR is based on a scale of 0–3: no dementia (CDR = 0), questionable dementia (CDR = 0.5), MCI (CDR = 1), moderate cognitive impairment (CDR = 2), and severe cognitive impairment (CDR = 3). Two sets of questions are asked, one for the informant and

another for the subject. The set for the informant includes questions about the subject's memory problem, judgment and problem solving ability of the subject, community affairs of the subject, home life and hobbies of the subject, and personal questions related to the subject. The set for subject includes memory-related questions, orientation-related questions, and questions about judgment and problem-solving ability.

E.3 Hamilton Depression Rating (HDR)

The Hamilton Depression Rating (HDR) is used to quantify the severity of symptoms of depression and is one of the most widely used and accepted instruments for assessing depression. The standard version of the HDR is designed to be administered by a trained clinician, and it contains 17 items rated on either a 3- or 5-point scale, with the sum of all items making up the total score. HDR scores are classified as normal (<8), mild depression (8 to 13), mild to moderate depression (14 to 16), and moderate to severe depression (>17). The HDR may be a useful scale for cognitively impaired patients who have difficultly with self-report instruments.

F Training Details

When training the coherence models, we sampled a new set of negatives (incoherent pairs of utterances) each time for a given narrative. Thus, after a few epochs, we covered the space of negative samples for even relatively long narratives. For discriminative models, we froze the sentence encoder after initialization to avoid overfitting. We run the models for 50 epochs with 4 epochs early stopping.

We used a grid search optimization technique to optimize the parameters. For consistency, we used the same experimental settings for all models. We first fine-tuned all models by performing a twenty-times grid search over their parameter pool. We empirically experimented with learning rate (lr): $lr \in \{0.00001, 0.00002, 0.00005, 0.0001, 0.0002\}$, batch size ($bs$): $bs \in \{16, 32, 64, 128\}$ and optimization (O): $O \in \{AdamW, Adam\}$. For the discrimination models, to tune the margin hyper-parameter (n), we experimented with the values $n \in \{3, 5, 7\}$. After the fine-tuning process, we trained again all the models for 50 epochs with 4 epochs, three times. We reported the average performance on the test set for all experiments. Model checkpoints were selected based on the minimum validation loss. Experiments were conducted on two GPUs, Nvidia V-100.