Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia

Dan Iter[1], Jong H. Yoon[2], and Dan Jurafsky[1]

[1]Department of Computer Science
[2]Department of Psychiatry and Behavioral Sciences
Stanford University
{daniter, jhyoon1, jurafsky}@stanford.edu

Abstract

Schizophrenia is a mental disorder which afflicts an estimated 0.7% of adults worldwide (Saha et al., 2005). It affects many areas of mental function, often evident from incoherent speech. Diagnosing schizophrenia relies on subjective judgments resulting in disagreements even among trained clinicians. Recent studies have proposed the use of natural language processing for diagnosis by drawing on automatically-extracted linguistic features, and particularly the use of discourse coherence. Here, we present the first benchmark comparison of previously proposed coherence models for detecting symptoms of schizophrenia and evaluate their performance on a new dataset of recorded interviews between subjects and clinicians. We also present two improved coherence metrics based on modern sentence embedding techniques that outperform the previous methods on our dataset. Finally, we propose a novel computational model for reference incoherence based on ambiguous pronoun usage and show that it is a highly predictive feature on our data. While the number of subjects is limited in this pilot study, our results suggest new directions for diagnosing common symptoms of schizophrenia.

1 Introduction

Schizophrenia is a severe mental disorder that affects thought, affective process and behavior. This paper focuses on one cardinal category of symptoms; *formal thought disorder*. *Formal thought disorder* (*FTD*) refers to disturbances in a person's thinking process, such as "flight of ideas" and distractibility (Andreasen, 1979). Symptoms of *FTD* can manifest as speech irregularities, generally perceived as a lack of coherence.

Psychiatrists diagnose schizophrenia by assessing subjects in a clinical setting and noting abnormalities based on the patient's reports of symptomology and their observed behavior. A reliable and automatic quantitative metric is desirable to effectively detect and treat schizophrenia. In other areas of medicine, metrics such as blood pressure or blood glucose levels are routinely used. However, no objective metrics of speech irregularities for schizophrenia are currently used in clinical settings. This pilot study extends the set of current academic models for detecting schizophrenia to further the development of such models for clinical use.

Recent academic literature has proposed measuring disorganized speech with semantic coherence, where larger amounts of concept overlap between two text segments is interpreted as more coherent (Bedi et al., 2015; Elvevåg et al., 2007). These proof-of-concept studies proposed using a coherence measure based on Latent Semantic Analysis (LSA) to quantitatively measure the presence or onset of *FTD* in subjects (Bedi et al., 2015; Elvevåg et al., 2007, 2010). In this pilot study, we present an empirical evaluation of these previous methods and a systematic comparison to our newly proposed methods for coherence.

We collected a new dataset of natural speech elicited by a formal interview with a trained clinician and evaluated the previously described methods for detecting symptoms of schizophrenia in text. We find that previously proposed methods are insufficient at modeling schizophrenia in our dataset. These methods incorrectly attribute greater coherence to longer sentences and greater use of verbal filler, problems that we suggest are fundamental to the class of algorithms using cosine similarity to model concept overlap. We introduce two new semantic coherence algorithms that correct for these systematic biases by leveraging recent advances in sentence and word embeddings to improving text representation. Both of

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 136–146
New Orleans, Louisiana, June 5, 2018.

these coherence models outperform the previously proposed methods and prove to be statistically significant discriminators between our schizophrenic and control groups.

We also investigate the use of referential incoherence in our schizophrenic groups. *FTD* has been reported to coincide with anomalies in deictic noun phrase usage, including various unusual uses of pronouns (Hinzen and Rosselló, 2015). We observed that referential incoherence, specifically the use of ambiguous pronouns, is a common pattern in incoherent speech. Ambiguous pronouns are pronouns whose reference is difficult for the listener to resolve because they refer to an entity that is never explicitly mentioned in the text, or one that is mentioned but only cataphorically, i.e., after the pronoun. Below is one example from the dataset where *they* is an ambiguous pronoun used to refer to the 49ers football team which is never mentioned:

> *Joe Montana* having a remarkable season coming off his Super Bowl Win where ***they*** upset the Cincinnati Bengals is off to another fabulous year

Figure 4 shows more examples of ambiguous pronoun use in our dataset. Based on this observation, we propose automatically measuring ambiguous pronoun usage as a novel computational model for referential incoherence in *FTD* and show its ability to predict schizophrenia in our pilot study.

2 Related Work

Speech analysis and coherence. *FTD* is typically diagnosed on the basis of the clinical observation of disorganized speech (Bedi et al., 2015; Adler et al., 1999). However, common clinical symptom assessment instruments or scales, such as Brief Psychiatric Rating Scale (BPRS) poorly capture many elements of *FTD* (Adler et al., 1999). There are other less commonly utilized clinical scales specifically established for measuring speech abnormalities, but many of these are hampered by the need for extensive and complex training for their proper administration or are based on subjective and non-quantifiable methods. This provides the primary motivation for using measures of coherence from natural language processing to quantify disorganized speech (Elvevåg et al., 2007).

Discourse coherence is the way parts of text are linked into a coherent whole, "a property of well-written texts that makes them easier to ... understand than a sequence of randomly strung sentences" (Lapata and Barzilay, 2005). Various aspects of discourse are associated with coherence. *Lexical cohesion* models chains of words and synonyms (Halliday and Hasan, 2014; Morris and Hirst, 1991). Relational models like *rhetorical structure theory* define *discourse relations* that hierarchically structure texts (Mann and Thompson, 1988; Lascaridest and Asher, 1991). *Referential coherence* focuses on the coherence of entities moving in and out of focus across a text (Grosz and Sidner, 1986; Barzilay and Lapata, 2008).

There are computational models of each of these aspects of coherence, but we focus here on lexical cohesion since it has attracted perhaps the most attention with relation to schizophrenia. LSA (Latent Semantic Analysis), the earliest dense vector embeddings models of word meaning, applies SVD (Singular value decomposition) to a matrix of word-document co-occurrences, and was applied early on as a model of discourse coherence, using cosines between embeddings for text regions as a measure of concept overlap or lexical cohesion (Foltz et al., 1998; McNamara et al., 2010).

Various other computational models have shown features of text and speech that can be automatically extracted and are associated with schizophrenia, including lexical features drawn from lexicons (Hong et al., 2012, 2015; Mitchell et al., 2015) and acoustic features (Covington et al., 2012). We focus in this paper on coherence metrics, but in the future will be exploring the role of these additional linguistic features on our dataset as well.

Models of coherence for schizophrenia. Elvevåg et al. (2007) were the first to propose computing coherence scores for predicting schizophrenia. They used LSA vectors to represent words in a text, ignoring syntax and treating each text as a bag-of-words, and compare texts by computing cosine similarities between their vector representations. From the beginning it was clear that this method relied on simplifying assumptions that might be inappropriate for schizophrenia; for example Foltz et al. (1998) notes that a discourse that simply repeated a sentence would be judged as highly coherent, problematic since repetition or perseveration can itself be a symptom of *FTD* (Andreasen, 1979; Hong et al., 2015).

There are two methods in the literature that have

used LSA embeddings and cosine similarity between representations to measure coherence for the purpose of detecting symptoms of schizophrenia. Confusingly they are both often referred to as "coherence" in the literature and so we will be assigning them distinct names, drawn from the terminology in describing *FTD* symptoms (Andreasen, 1979).

What we will call the *Tangentiality Model* (Elvevåg et al., 2007) uses the coherence metric to compare fixed-sized word windows of responses to their corresponding questions. The coherence of a response is computed as the slope of the linear regression line for the cosine similarities of the sliding window. Steeper slopes mean the response is moving further away from the question and therefore becoming more incoherent.

What we will call the *Incoherence Model* (Bedi et al., 2015) measures the coherence of a speaker by computing the semantic coherence of each adjacent pair of sentences in a document to derive a global coherence independent of the question, which they call First Order Coherence. Bedi et al. (2015) choose to use the minimum coherence score per document as a feature in a convex hull classifier for predicting schizophrenia. Thus the methods differ in whether the speaker's text is compared to the speaker's prior text, or to the interviewer's question. As we will see, both of these naive embedding-based coherence metrics have problems at detecting *FTD* on conversational dialog.

Ambiguous pronouns. To our knowledge, there have been no previous efforts to automatically measure ambiguous pronoun use as a feature of schizophrenia. Novogrodsky and Edelson (2016) reports increased ambiguous pronoun usage, including cataphora, among children with Autism Spectrum Disorder. Hinzen and Rosselló (2015) notes "pronouns are often used without their reference being clear to the listener, and they fail to track referents across discourse" which implies that measuring untracked references may provide a strong predictive signal.

Schizophrenia datasets. A challenge for computational linguistics efforts in schizophrenia is the dearth of publicly available patient data. This motivated us to collect our own data of naturalistic speech spoken by individuals with schizophrenia. Previous studies have used datasets ranging from 5-23 schizophrenics and similar numbers of controls (Bedi et al., 2015; Elvevåg et al., 2007; Hong et al., 2015). Our pilot study has 5 controls and 9 schizophrenic patients which is similar in size to these studies. Mitchell et al. (2015) used text from social media by self-reporting schizophrenics which is much larger but does not contain any psychiatric assessments. Some studies explore a large number of features over relatively small datasets, thus increasing the likelihood of a multiple comparisons problem (Hong et al., 2015). In our pilot study, we also operate on a small dataset but attempt to analyze failure cases to support intuitions as to how each method may generalize.

3 Dataset

Stat	Total	SZ	Control
Words	37,673	29,103	8,570
Sentences	2,272	1,824	448
Responses	123	82	41
Avg Resp/ Subject	8.78	9.11	8.2
Avg Words/ Resp	306.28	354.91	209.02

Table 1: Summary statistics for collected interview transcripts. Note that each response is relatively long making the interview a series of extensive responses to short prompts.

We evaluate our models on a new dataset collected from subjects with schizophrenia or a closely related condition, schizoaffective disorder and from psychiatrically healthy comparison subjects. Patients were recruited from in patient and outpatient psychiatric services. Control subjects were recruited from the local community. Experienced doctoral level clinicians confirmed the diagnoses of schizophrenia or schizoaffecive disorder in patients and the absence of major psychiatric conditions in control subjects using the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) criteria. Patients' symptoms were characterized with standard clinical instruments, including the Scale for the Assessment of Positive Symptoms (SAPS) and the Scale for the Assessment of Negative Symptoms (SANS), assessed by the second author, a psychiatrist with many years of administering these instruments. After a complete description of the study was provided, written informed consent was obtained from study participants. The study was approved by the Institutional Review Board of Stanford University.

The dataset consists of interviews with 14 subjects, 5 controls (free of major psychiatric illness)

and 9 patients diagnosed with schizophrenia or schizoaffective disorder. The control and interest groups consist of 100% and 80% males with mean ages of 40.3 and 29.5, respectively. Each interview consists of a 15-30 minute one-on-one interview with research staff that asks them 8-10 questions, such as "describe your favorite book or movie", "describe something interesting that you did recently", "describe the room we are in". The full set of questions can be found in the appendix. Table 1 contains some high level statistics about the dataset. It is worth noting that the average length of a response is about 300 words, while the questions are relatively short. Therefore, we analyze the text not as multi-turn dialogue but rather as a collection of monologues that are prompted by the interview questions. This motivates our decision to segment the data into question and response pairs as well as to analyze the responses on the sentence granularity rather than utterance or turn. The interviews were recorded with high-quality digital stationary room microphones and transcribed by a professional transcription service, using standard linguistic conventions, marking all the words spoken, and assigning time markings to allow us to align the transcribed text exactly with the acoustics.

We use only the text transcripts in this analysis, ignoring for the moment acoustic features such as pitch, energy and rate of speech, although we plan to investigate these in future work. We do some minor preprocessing on the transcripts to group the responses per question and backchannels (e.g., *OK*, *uh-huh*) from the interviewer during the response. However, we keep all transcribed details of the response, including filled pauses, word fragments, mispronunciations and repetitions.

4 Coherence Models for Schizophrenia

Formal thought disorder is typically diagnosed on the basis of a clinical observation of incoherent speech (Bedi et al., 2015). An example of incoherence in our dataset follows:

> "When I was three years old, I made my first escape attempt. I had a [unintelligible] sticker in the window. Like everybody listened to AM radio in the sixties. They had a garage band down the street. I couldn't understand why the shoes were up on the wire. That means there was drug deal in the neighborhood."

The above example is an instance of derailment, a symptom of *FTD*, where there is little semantic overlap between sentences (Andreasen, 1979). The characteristic of unrelated sentences is a motivation for using LSA-based semantic overlap to measures coherence. This section outlines two prominent models for coherence in the domain of schizophrenia, provides an analysis of failure cases for these baselines and presents our improvements to the current state-of-the-art.

4.1 Baseline Coherence Models

Currently, there are two reported methods for measuring coherence in the context of schizophrenia, both of which model coherence as the concept overlap between two texts (Bedi et al., 2015; Elvevåg et al., 2007). We evaluate both of these methods as the baselines in this study, and show how to update both of these methods with our proposed improvements.

For both models, each sentence or window of tokens is embedded by taking the average of word vectors generated from LSA word embeddings, and both models train the LSA word embeddings on the Touchstone Applied Science Associates (TASA) Corpus of school texts with a mix of age-graded reading levels (Bedi et al., 2015; Elvevåg et al., 2007). There are two different models for measuring coherence using this representation. As discussed above, since both models are confusingly referred to as *coherence* in the literature, we give them separate names.

The *Incoherence Model* (named after the Andreasen (1979) definition of "Incoherence" focusing on unintelligible combinations of words) is computed by scoring each adjacent pair of sentences in a subject response by the cosine similarity between the two sentence embeddings (Equation 1). The coherence of a response (or document) is the mean of all the cosine similarities and the coherence of a subject is the mean of the scores for all responses.

For the *Tangentiality Model* (named after the Andreasen (1979) definition of "Tangentiality" where a speaker wanders from a topic and never returns), a linear regression line is fit to cosine similarities between the interviewer's question and a moving fixed-sized window of the subject's response. The slope of the regression line is the co-

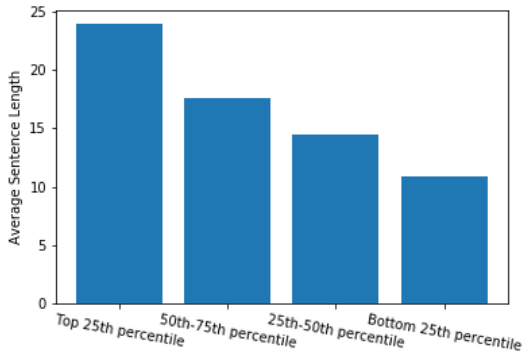

Figure 1: Average length of a sentence in each quartile of coherence scores.

herence metric. A steeper slope indicates the response is moving further away from the question and therefore is less coherent.

$$similarity = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}|| \quad ||\vec{B}||} \tag{1}$$

4.2 Error Analysis

We implement and evaluate the above described methods as our baselines.[1] The two baselines algorithms are **not** able to significantly capture the difference between schizophrenics and controls in our dataset. Figure 2 shows the *Incoherence Model* scores for each subject. This baseline metric does not significantly distinguish between the two groups (t-test statistic = 0.487, p = 0.634). We found three primary failure cases that add noise to both coherence metrics; (1) verbal filler, (2) bias toward longer sentences and (3) repetition.

Table 2 shows the 10 least coherent sentence pairs as scored by the baseline *Incoherence Model*. Many of the examples contain filled pauses, such as "um". Filled pauses are enormously common in conversational speech and not not generally considered a sign of incoherence, and furthermore and there is no evidence to suggest they are a symptom of schizophrenia. This seems a problem with the baseline algorithms.

Figure 1 shows that the top 25th percentile of sentence pairs have an average sentence length of 24 words while the bottom 25th percentile of sentence pairs have an average of less than 11 words. Elvevåg et al. (2007) alludes to this issue, noting that their metric assigned higher coherence scores with longer windows, since the coherence (cosine)

[1]We use SpaCy's sentence tokenizer and extract question-response pairs manually.

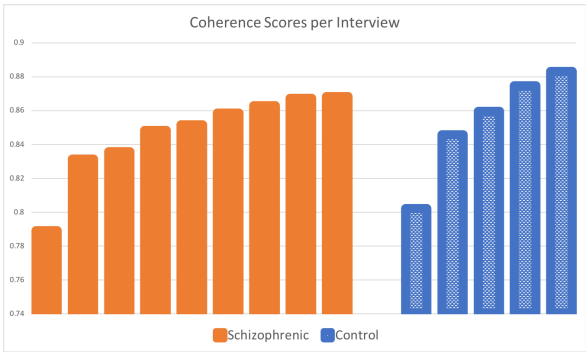

Figure 2: **Baseline Incoherence Metric**: Each bar is the coherence score for one subject computed as the mean of the cosine similarities of all adjacent sentences in a response. Each sentence is embedded as the mean of the word vectors.

typically increases with a bigger window size due to greater contextual overlap (i.e., more similar words).

Finally, there are some sentences with significant repetition, which is in fact a symptom of thought disorder; perseveration (Andreasen, 1979; Hong et al., 2015). Since a coherence metric treats a sentence as a bag of words and measures the overlap, repeated words can result in sentence pairs being scored as highly coherent when they are completely unintelligible. This can be seen in an extreme case, where a single word is repeated for the entire discourse.

For example the following excerpt from the dataset is scored as highly coherent by the baseline *Incoherence Model* (0.981) but is in fact not extremely coherent to a human reader:

> "Like he'll make me feel he'll take away my laptop and be like if you ever, you want to *steal*, this is what it feels like to be, to have your *stuff stolen* and he'll take it just temporarily, you know, just to make it, me feel like what it's like to, to have my *stuff stolen*. He's like do you really want to go around, you know, making other people feel like the other *stuff's stolen*, you know?"

4.3 New Coherence Models

The challenges outlined in Section 4.2 are fundamental to the class of algorithms using cosine similarities of embeddings as a measure of concept overlap. Therefore, we apply identical improvements to the *Tangentiality Model* and the *Incoherence Model* to produce two new algorithms

Uhm.	Narrative meaning?	0.406
Woo.	A little ball hitting the other ball	0.387
Um,	but	0.380
Hexagonal?	I don't know the name.	0.355
It's something else.	Hexagonal?	0.350
Or	yeah.	0.332
Uh let me think of one first.	Um	0.323
Um	So, all right.	0.284
Um, I guess it's a vacation as opposed a trip then.	Um, badum badum.	0.218
Um, badum badum.	A vacation.	0.184

Table 2: The 10 lowest scoring pairs of sentences in our corpus. Less coherent pairs have lower scores.

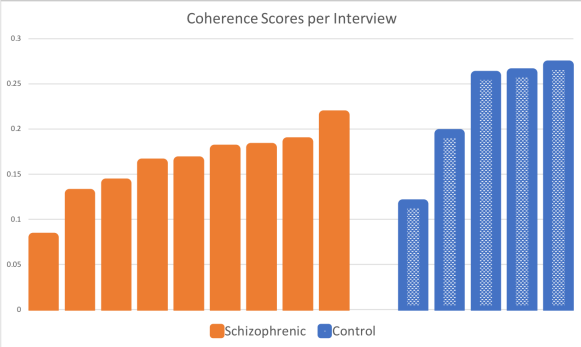

Figure 3: **New Incoherence Metric** Each bar is the coherence for one subject using the improvements explained in Section 4.3. Verbal filler and sentences entirely composed of stop words are removed. Words are embedded with Word2Vec and sentences are embedded with SIF sentence embedding.

that measure the same specific forms of coherence. Our two key innovations are (1) preprocessing the data to deal with conversational characteristics and (2) employing modern word and sentence embeddings to improve the representation. We show that by applying these improvements to both baselines, the resulting algorithms differentiate between our two subject groups with statistical significance and are strong predictors of schizophrenia.

The preprocessing changes are simple. First, we remove all filler words (i.e., various forms of *uh*, *um*, *you know*, etc.) and sentences entirely composed of stop words. Second, we replace the sliding window in the *Tangentiality Model* with sentence tokenization to capture semantically meaningful chunks of the response, obviating the need to tune the window size parameter.

Second, we draw on recent advances in word embeddings (Mikolov et al., 2013; Pennington et al., 2014) and sentence embeddings (Arora et al., 2016; Pagliardini et al., 2018). These are known to provide superior representations, such as correcting for sentence embeddings that contain "semantically meaningless directions" (Arora et al., 2016). We test a number of sentence embeddings, which we refer to as TF-IDF (Lintean et al., 2010), Sent2Vec (Pagliardini et al., 2018) and Smooth Inverse Frequency (SIF) (Arora et al., 2016).

TF-IDF is a traditional vector weighting scheme; in using it to create sentence embeddings we follow the parameterization of Lintean et al. (2010), proposed originally to create sentence embeddings for LSA: multiplying each word embedding by the raw (non-logged) term frequency (# of times that word occurs in the sentence) and dividing by the (non-logged) document frequency (# of documents in which the term is used in a corpus). Typically, for small corpora the denominator term is taken from a large corpus; we chose the enwiki dataset (Wikimedia, 2012). Sent2Vec learns a new word embedding similar to Word2Vec but extends the training objective such that each sentence embedding is predictive of the sentences around it. Intuitively, common words would be less predictive of the surrounding sentences and therefore should play a smaller role in the embedded sentence representation. SIF also computes a weighted average of each sentence, similar to TF-IDF, followed by removal of the projection of the first principal component of the singular value decomposition of the sentence embedding matrix. This common component removal is expected to remove the "semantically meaningless direction" described by Arora et al. (2016) that may be captured by common terms in the dataset that may not be common in general.

The three sentence embedding techniques mentioned above are intended to improve the em-

bedded representation for sentences. They each take different approaches to removing semantically meaningless terms from the representation. The intuition here is that the bias toward longer sentences in the baseline coherence metric is due to the large overlap of semantically meaningless words (such as stop words) which can be removed with smooth inverse frequency or weighted averaging of terms by term frequency. TF-IDF, SIF and Sent2Vec all correct for this meaningless word and longer sentence bias.

Table 3 shows SIF and Sent2Vec sentence embeddings both outperforming mean vector sentence embedding (used in the baseline models) in significantly distinguishing between the two subject groups for both the *Incoherence Model* and the *Tangentiality Model*. Interestingly, while TF-IDF term weighting often fall in between mean vector and SIF in terms of the t-test statistic for the *Incoherence Model*, it performs well for the *Tangentiality Model* using LSA word embeddings and more poorly for the other embeddings. However, TF-IDF is still outperformed by SIF using both Glove and Word2Vec word embeddings. Our improvements to both coherence models are sufficient to assign significantly higher coherence to our control subjects than our schizophrenic subjects. Figure 3 shows the coherence scores output by our new *Incoherence Model*.

Note that our improvements do not yet address the issue of word repetition. Since repetition itself is a symptom of schizophrenia (Hong et al., 2015), we need a more powerful model of what constitutes abnormal repetition as opposed to natural lexical cohesion, presumably a model that will need to draw on other linguistic markers.

Distinguishing Schizophrenics from Controls		
Incoherence Model		
Sentence	Word	t-test Stat
Mean Vector	LSA	0.594
	Glove	0.514
	Word2Vec	1.147
TD-IDF	LSA	1.142
	Glove	0.935
	Word2Vec	1.957
SIF	LSA	1.517
	Glove	2.139
	Word2Vec	2.432*
Sent2Vec	Sent2Vec	2.067
Tangentiality Model		
Sentence	Word	t-test Stat
Mean Vector	LSA	0.588
	Glove	1.820
	Word2Vec	1.689
TF-IDF	LSA	2.173*
	Glove	0.718
	Word2Vec	1.372
SIF	LSA	1.930
	Glove	2.207*
	Word2Vec	2.353*
Sent2Vec	Sent2Vec	1.085

Table 3: Population difference between positive and control subjects measured by t-test using different word and sentence embeddings for two coherence metrics. (*) signifies statistically significant with p-value less than 0.05. See appendix for full table containing p-values, means and standard deviations for each group.

5 Referential Coherence Model

We next propose a novel model for measuring coherence, ambiguous pronoun usage, based on earlier work pointing out referential problems in schizophrenics (Hinzen and Rosselló, 2015). *Ambiguous pronoun usage* is the reference to an entity using a pronoun that is either (1) never resolved or (2) resolved after the use of a proper noun (cataphora). Figure 4 shows samples from our dataset, including examples of cataphora that create notable confusion in the sentence. We present the following algorithm to automatically measure the number of ambiguous pronouns used by subjects during clinical assessments:

1. Co-references are extracted from the corpus with a pretrained co-reference resolver (Lee et al., 2017).

2. For each document, for each entity, the model outputs a reference chain (a list of terms that should refer to the same entity.)

3. The ambiguous pronoun count for each subject is the total number of cases where the first term in a list of entity references is a third-person pronoun (he, she, they, etc.).

All but one schizophrenic subject in our study exhibited at least one case of ambiguous pronoun use and on average 3.2 cases. Two controls have zero cases of ambiguous pronoun use and there is exactly one case in each of the other 3 controls. The most common ambiguous pronoun used was *they* followed by *he* and *them*. All ambiguous pro-

(1) Well it's a ... I believe ***they*** use it, it's a multi-purpose room. ***They*** use it for report, ***they*** have snacks in here, ***they*** interview patients.

(2) *Joe Montana* having a remarkable season coming off his Super Bowl Win where ***they*** upset the Cincinnati Bengals is off to another fabulous year

(3) I always pour water over sands and where ***he*** would hold, my, ***my brother***, [Samuel], would be serving the mass with me. And he would hold the bowl so the water wouldn't get on the carpeting.

(4) Sure, I had fun... and I'd scream at ***him***, like a girl, so ***[Dalton]*** says.

Figure 4: Above are examples of ambiguous pronoun usage from our dataset. Personal names in square brackets were changed for anonymity. (1) the speaker refers to a third person entity that is never named. (2) *they* refers to Joe Montana's team, but the team is never named. Resolving *they* to refer to Joe would mean the incorrect pronoun is used. (3) and (4) are both cases of cataphora. Pronouns are bold. Candidate entities are underlined. Bold and underlined entities are correctly resolved. Dotted lines indicate incorrect resolution. Missing lines from a pronoun indicate ambiguity.

nouns were third person, 19 were plural and 11 were singular. Figure 5 shows the total counts for each subject.

Because the scores are generated automatically using a co-reference resolution tool that is trained on written text rather than transcribed speech, the signal is noisy due to errors in the resolutions. Nonetheless, the fact that ambiguous pronouns are detected significantly more often among schizophrenics suggests that there is a deviation in the speech patterns that this metric is identifying.

Finally, to underscore the predictive power of this model as a marker of clinical symptology, we show that ambiguous pronoun usage counts strongly correlate with a number of clinical metrics in our dataset. The Spearman correlation coefficients of correlations for ambiguous pronoun usage with Global Thought Disorder is 0.749 and with Scale for the Assessment of Negative Symptoms (SANS) is 0.732, both of which correlate with p-values less than 0.01.

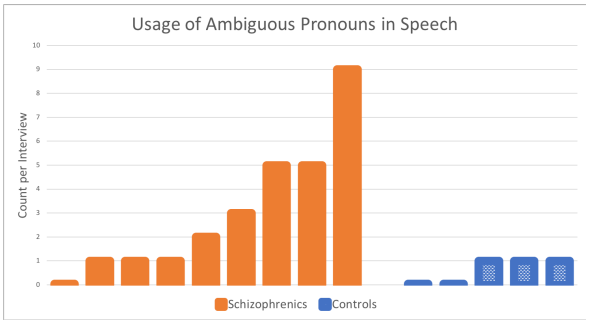

Figure 5: Ambiguous pronoun usage scores for all subjects using automatic co-reference resolution.

6 Classification

We train a classifier to show the predictive power of the features we discussed and the relative importance of each feature. Due to the small size of our dataset, we make no claim to the generalization of this classifier on new data. Furthermore, we report the feature importance scores to give some notion of their relative effects, though their significance and generalizability is limited due to the small training data set size. A Random Forest binary classifier is able to achieve 93% accuracy and Logistic Regression achieves 86% accuracy in separating the control and schizophrenic groups with leave-one-out cross validation. Logistic regression was trained with L2 regularization (C=0.01) and the Random Forest classifier was trained with 10 estimators, using 1 feature at each split with a max depth of 5. All parameters were chosen using grid search. We report both because Random Forests are often effective in linguistic tasks while Logistic Regression is often used for feature importance analysis. We use only three features: both coherence measures (using the best embeddings) and ambiguous pronoun counts. Table 4 contains the feature importance for the Random Forest classifier and coefficients from Logistic Regression. Logistic Regression misclassifies two schizophrenics. Both misclassified subjects are somewhat anomalous in that they had only one case of ambiguous pronoun usage each and relatively high coherence scores.

7 Conclusion

In this pilot study, we explore two linguistic phenomena: coherence measured using concept overlap, and ambiguous pronoun usage, as features for objectively measuring *FTD*. We show that previous methods for measuring coherence may fail to

Feature	RandForest	LogReg
Incoherence Model	0.443	-0.058
Tangentiality Model	0.363	-0.048
Ambiguous Pronouns	0.188	0.044

Table 4: Feature importance scores from Random Forest classifier with 93% accuracy and Logistic Regression with 86% accuracy with leave-one-out cross validation. Scores reported are coefficients from the Logistic Regression model and the feature importance attributes of the Random Forest model. Both quantities are attributes of the respective SciKit Learn objects.

be representative of the underlying text because of common biases due to common words and sentence length. and describe two improvements: filtering verbal fillers and sentences composed entirely of stop words, and employing modern word and sentence embeddings to improve text representation. In particular, we show that the modern word and sentence embeddings outperform LSA-based word embeddings with both mean vector and TF-IDF weighted sentence embeddings on our dataset. Finally, we present a novel computational feature for referential coherence based on ambiguous pronouns.

On our new dataset, these computational features significantly distinguish between subjects with schizophrenia and controls, and correlate strongly with clinical ratings that are commonly used for assessing patients, and improve over strong baselines. We also introduce a classifier that is able to achieve 93% accuracy on our dataset with leave-one-out cross-validation. We present these findings to further the study of reliable and objective metrics of *FTD* among schizophrenics for the purpose of clinical assessment.

Acknowledgments

This research was partially funded by the NSF via grant IIS-1514268.

References

Caleb M Adler, Anil K Malhotra, Igor Elman, Terry Goldberg, Michael Egan, David Pickar, and Alan Breier. 1999. Comparison of ketamine-induced thought disorder in healthy volunteers and thought disorder in schizophrenia. *American Journal of Psychiatry*, 156(10):1646–1649.

Nancy C Andreasen. 1979. Thought, language, and communication disorders: Ii. diagnostic significance. *Archives of general Psychiatry*, 36(12):1325–1330.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. *International Conference on Learning Representations (ICLR)*.

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

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*, 1:15030.

Michael A. Covington, SL Anya Lunden, Sarah L. Cristofaro, Claire Ramsay Wan, C. Thomas Bailey, Beth Broussard, Robert Fogarty, Stephanie Johnson, Shayi Zhang, and Michael T. Compton. 2012. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophrenia research*, 142(1):93–95.

Brita Elvevåg, Peter W Foltz, Mark Rosenstein, and Lynn E DeLisi. 2010. An automated method to analyze language use in patients with schizophrenia and their first-degree relatives. *Journal of neurolinguistics*, 23(3):270–284.

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):304–316.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics. Formerly the American Journal of Computational Linguistics*, 12(3).

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. Routledge.

Wolfram Hinzen and Joana Rosselló. 2015. The linguistics of schizophrenia: thought disturbance as language pathology across positive symptoms. *Frontiers in psychology*, 6:971.

Kai Hong, Christian G. Kohler, Mary E. March, Amber A. Parker, and Ani Nenkova. 2012. Lexical differences in autobiographical narratives from schizophrenic patients and healthy controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 37–47. Association for Computational Linguistics.

Kai Hong, Ani Nenkova, Mary E March, Amber P Parker, Ragini Verma, and Christian G Kohler. 2015. Lexical use in emotional autobiographical narratives of persons with schizophrenia and healthy controls. *Psychiatry research*, 225(1):40–49.

Mirella Lapata and Regina Barzilay. 2005. Automatic evaluation of text coherence: Models and representations. In *19th International Joint Conference on AI*, volume 5, pages 1085–1090.

Alex Lascaridest and Nicholas Asher. 1991. Discourse relations and defeasible knowledge‘. In *29th Annual Meeting of the Association for ational Linguistics*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197. Association for Computational Linguistics.

Mihai Lintean, Cristian Moldovan, Vasile Rus, and Danielle McNamara. 2010. The role of local and global weighting in assessing the semantic similarity of texts using latent semantic analysis. *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Danielle S. McNamara, Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

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

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 11–20.

Jane Morris and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1).

Rama Novogrodsky and Lisa R Edelson. 2016. Ambiguous pronoun use in narratives of children with autism spectrum disorders. *Child Language Teaching and Therapy*, 32(2):241–252.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised learning of sentence embeddings using compositional n-gram features. *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sukanta Saha, David Chant, Joy Welham, and John McGrath. 2005. A systematic review of the prevalence of schizophrenia. *PLoS medicine*, 2(5):e141.

Wikimedia. 2012. English wikipedia dump. Http://dumps.wikimedia.org/enwiki/latest/enwiki-latestpages-articles.xml.bz2.

A Supplemental Material

A.1 Interview Questions

- Could you please tell me about your favorite book, TV show, video game, or board game. Please pretend that I've never heard of this book or show or video game or board game so I that I can understand.
- Could you please describe your favorite childhood memory?
- Could you tell me about your favorite hobby and how one does it?
- What's an interesting thing you've done or seen recently? Why did you find interesting?
- Could you tell me about a typical day for you?
- Could you tell me how you brush your teeth?
- Could you please give me a detailed description of the room we are in?
- Could you please tell me about the most memorable recent day you had?
- Could you please tell me about your best friend?
- Could you please tell me about your relationship with your mother?
- Could you tell me about the community or neighborhood you live in?
- Could you give me a detailed description of the chair you're sitting in?
- Could you tell me about a trip you've taken at some point. It could be any time in your life.
- Could you tell me how one searches for something on the internet?
- Could you tell me how you would go about making a sandwich?

A.2 Extended experimental results

Distinguishing Schizophrenics from Controls							
Incoherence Model							
Sentence	Word	t-test Stat	p-value	SZ Mean	SZ Std	Control Mean	Control Std
Mean Vector	LSA	0.594	0.563	0.312	0.044	0.328	0.049
	Glove	0.514	0.616	0.846	0.022	0.853	0.028
	Word2Vec	1.147	0.272	0.628	0.032	0.653	0.046
TD-IDF	LSA	1.142	0.274	0.323	0.048	0.355	0.043
	Glove	0.935	0.367	0.438	0.023	0.454	0.039
	Word2Vec	1.957	0.072	0.319	0.039	0.364	0.040
SIF	LSA	1.517	0.153	0.114	0.024	0.134	0.022
	Glove	2.139	0.052	0.182	0.059	0.278	0.103
	Word2Vec	2.432*	**0.030**	0.151	0.044	0.221	0.059
Sent2Vec	Sent2Vec	2.067	0.059	0.235	0.043	0.285	0.038
Tangentiality Model							
Sentence	Word	t-test Stat	p-value	SZ Mean	SZ Std	Control Mean	Control Std
Mean Vector	LSA	0.588	0.567	1.39e-4	5.59e-4	4.52e-4	1.36e-3
	Glove	1.820	0.092	-9.27e-5	3.29e-4	2.37e-4	2.62e-4
	Word2Vec	1.689	0.115	-3.55e-4	2.87e-4	-3.64e-6	4.58e-4
TF-IDF	LSA	2.173*	**0.049**	-2.46e-4	5.26e-4	7.30e-4	1.09e-3
	Glove	0.718	0.485	-6.02e-4	2.09e-3	1.47e-4	8.11e-3
	Word2Vec	1.372	0.193	-7.96e-4	1.89e-3	7.07e-4	1.80e-3
SIF	LSA	1.930	0.076	-1.80e-4	5.36e-4	1.19e-3	1.95e-3
	Glove	2.207*	**0.046**	-2.89e-5	7.94e-4	1.05e-3	8.99e-4
	Word2Vec	2.353*	**0.035**	-1.73e-4	7.33e-4	9.59e-4	9.66e-4
Sent2Vec	Sent2Vec	1.085	0.298	9.538e-5	3.03e-4	1.16e-4	3.83e-4

Table 5: Population difference between positive and control subjects measured by t-test using different word and sentence embeddings for two coherence metrics. (*) signifies statistically significant with p-value less than 0.05.