# Are Interaction Patterns Helpful for Task-Agnostic Dementia Detection? An Empirical Exploration

**Shahla Farzana** and **Natalie Parde**
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois Chicago
{sfarza3, parde}@uic.edu

## Abstract

Dementia often manifests in dialog through specific behaviors such as requesting clarification, communicating repetitive ideas, and stalling, prompting conversational partners to probe or otherwise attempt to elicit information. Dialog act (DA) sequences can have predictive power for dementia detection through their potential to capture these meaningful interaction patterns. However, most existing work in this space relies on content-dependent features, raising questions about their generalizability beyond small reference sets or across different cognitive tasks. In this paper, we adapt an existing DA annotation scheme for two different cognitive tasks present in a popular dementia detection dataset. We show that a DA tagging model leveraging neural sentence embeddings and other information from previous utterances and speaker tags achieves strong performance for both tasks. We also propose content-free interaction features and show that they yield high utility in distinguishing dementia and control subjects across different tasks. Our study provides a step toward better understanding how interaction patterns in spontaneous dialog affect cognitive modeling across different tasks, which carries implications for the design of non-invasive and low-cost cognitive health monitoring tools for use at scale.

## 1 Introduction

A recent surge of interest in automated assessment of cognitive health within the speech and language processing communities (Zhu et al., 2019; Di Palo and Parde, 2019; Farzana and Parde, 2020; Luz et al., 2020, 2021) has spurred the development of high-performing diagnostic models. These models carry the potential for substantial real-world positive impact, offering an affordable and accessible healthcare screening solution for individuals who may otherwise be under-served (Petti et al., 2020). However, recent cognitive assessment models have generally been constrained to specific tasks, each with their own characteristics and requirements. Although this facilitates the development of models that excel at their target task (e.g., predicting which users have Alzheimer's disease in a picture description task (Luz et al., 2020)), it creates challenges in building generalizable knowledge about the complex relationship between linguistic or verbal behavior and cognitive status. It can be unclear which findings are task-specific, and which may be applicable to different tasks in related settings.

In this work, we set out to provide clarity regarding the generalizability of a category of features that have held promise for task-specific cognitive assessment. Specifically, we examine facets of individuals' interaction patterns, which have been recognized as predictive of Alzheimer's disease or related dementia (AD) in sociolinguistic studies (Orange et al., 1996; Elsey et al., 2015; Hamilton, 1994) and proved informative for automatically detecting AD in task-specific settings (Nasreen et al., 2021; Mirheidari et al., 2019). We do so by adapting an existing dialog act (DA) annotation scheme (Bunt, 2006; Farzana et al., 2020), previously used to analyze dialogs from AD and control participants, to two distinct cognitive tasks and study subjects' interactions across tasks. We also examine the use of interaction features derived from these DA tags in dementia detection models to assess their task-agnostic utility in this domain. Our key contributions are as follows:

- We adapt a DA annotation scheme for two cognitive tasks in a popular dementia detection corpus and present comparative analyses of subjects' interaction patterns across tasks.
- We develop a DA tagging model using this scheme and show that it achieves strong performance ($F_1$=0.82) when trained on both tasks jointly. The model leverages neural sentence embeddings, part-of-speech (POS) tags, previous utterances, and speaker information

to make its predictions.

- We propose a set of content-free interaction features for task-agnostic dementia detection and show that they yield high utility in distinguishing between dementia and control subjects across different tasks.

We describe these contributions further in the remainder of this paper. In §2, we review relevant background to position our work within the broader research landscape. In §3, we present our methods for modeling DA sequences using the selected DA scheme (§3.1) and developing task-agnostic interaction features within this domain (§3.2). We describe our data in §4, our experiments in §5, and our results in §6, before concluding in §7.

## 2 Background

### 2.1 Interaction Patterns and AD Detection

Conversation analysis has proved to be effective for detecting dementia and tracking its progression through the study of user intent, clarification and verbal disfluency frequency, and other discourse cues (Mirheidari et al., 2019; Orange et al., 1996; Farzana et al., 2022). Speech-based interaction features like average turn duration, total turn duration, and average number of words per minute have been utilized to model conversation dynamics in the context of AD detection (Luz et al., 2020). However, many of these features are task-specific, and focus only on the participants' part of the dialog. Nonetheless, fine-grained analysis of question-answer ratio has been the focus of several studies showing promising performance on dementia detection (Hamilton, 1994; Varela Suárez, 2018).

Dialog act-based conversation analysis was first introduced by Farzana et al. (2020), capturing the interaction patterns from DementiaBank's (Becker et al., 1994) semi-structured picture description task in terms of different DAs from both the subject and interviewer. Similar corpus analyses on the Carolinas Conversation Collection (CCC) (Pope and Davis, 2011) by Nasreen et al. (2019) observed that interaction patterns like signal non-understanding and clarifying questions are more evident in cognitively challenged subjects than healthy controls, and leveraged DA features to model dementia detection (Nasreen et al., 2021). Speaker turn sequence processing has also previously been used to model intervention patterns (Sarawgi et al., 2020), and leveraging acoustic, linguistic, and fusion features to represent conversations between interviewers and participants has shown promising performance in AD detection (Pérez-Toro et al., 2021). Most of these experiments have been evaluated on task-specific corpora, including semi-structured cognitive screening interviews like the picture description task (Roth, 2011) or more open-ended tasks in which subjects talk about their health (Pope and Davis, 2011)). Modeling task-agnostic linguistic anomalies to detect dementia from casual conversations has been studied very recently (Li et al., 2022), although this work did not extend its study to interaction style.

### 2.2 DA Tagging and AD Detection

DA recognition is known to be a complex problem, and many approaches ranging from multi-class/multilabel classification to structured prediction have sought to tackle it (Stolcke et al., 2000; Yang et al., 2009). Performing DA classification effectively enables the development of high-quality natural language dialogue systems (Higashinaka et al., 2014). Previously, a context-aware deep neural model leveraging a hierarchical recurrent network and self attention mechanism (Raheja and Tetreault, 2019) achieved state-of-the-art performance in DA tagging on the SWDA corpus (Jurafsky et al., 1997), a standard benchmark for this task.

Most DA tagging corpora are highly imbalanced, so a crucial shortcoming of most high-performing DA tagging models is that in focusing on improving overall performance, they end up performing poorly on rare class DAs. These DA classes can be critical for modeling conversations in cognitive health screening tasks (Farzana et al., 2020; Nasreen et al., 2021). Thus, DA tagging models tailored more specifically for AD detection settings may be needed to facilitate sufficient understanding and analysis of interaction patterns.

## 3 Methods

### 3.1 DA Tagging Model

Our initial dialogue act recognition model trained on Farzana et al. (2020)'s *Cookie-Theft DA* dataset is a multi-layer perceptron (MLP) adapted from a model introduced in prior work (Martínek et al., 2021). Each utterance, consisting of a variable number of words, is first encoded into a single pre-trained 1024-dimensional sentence embedding vector using a BERT Large (Reimers and Gurevych,

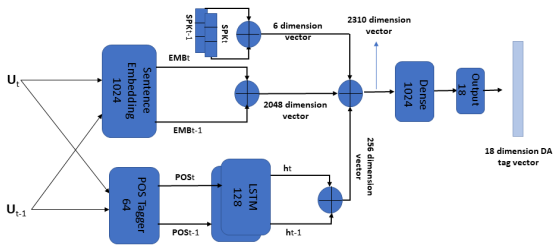


Figure 1: DA Tagging model architecture.

2019) encoder. As shown in Figure 1, our model computes two such vectors, respectively, for the context and current utterances. These vectors are concatenated and passed along as input to the MLP. We also incorporate utterance-wise part-of-speech (POS) tags generated using the pre-trained Stanford CoreNLP parser (Qi et al., 2020). To do so, we feed sequences of POS tags for the current and contextual utterance through an LSTM and concatenate its output with the previously computed semantic representation as shown in Figure 1.

We also add a speaker information vector indicating the speaker for a given utterance. We compute one speaker vector each for the current and context utterances and concatenate them with the previously created representation, ultimately resulting in a concatenation of numerous input vectors (utterances, POS sequences, and speaker tags) that is fed to the dense layer of the MLP, followed by the output layer. Following the training procedures later described in §5.1, we perform DA tagging experiments on two AD detection tasks separately and in a joint setting. In doing so, we seek to investigate the following topics pertaining to DA prediction: (1) the model's ability to generalize when predicting DAs for two different cognitive tasks, provided that the tasks share some common nuances in interaction style, yet differ in linguistic traits and overall objectives; and (2) the extent to (and ways in) which prediction accuracy for rare DAs differs when the model is trained jointly with a class-weighted loss function versus when the model is trained separately on single tasks.

### 3.2 AD Detection Features

To investigate the effects of interaction patterns on AD detection performance in numerous cognitive tasks, we made use of the DA tags as well as turn-based features, following their earlier success in prior work (Nasreen et al., 2021). We represented local interaction patterns as unigram, bigram, and trigram sequences of DA tags. To avoid sparsity, we filtered these n-grams based on their training set frequency differences between the AD and non-AD classes (i.e., only DA n-grams for which the between-class training set frequency differed by $\geq 5$ were retained). To represent turn-taking patterns, we computed the following based on timing signatures from the transcripts:

- **Average Turn Duration:** The average length of a participant's turn, in milliseconds.
- **Total Duration:** The length of the full conversation (in milliseconds) between the participant and interviewer.
- **Normalized Turn Switch:** The average number of turn switches per minute (e.g., a minute of dialog with the turn sequence (*Participant* → *Interviewer* → *Interviewer* → *Participant*) would have two turn switches).
- **Average Words/Minute:** The average number of words spoken in a minute of recorded speech.

These features may provide valuable clues regarding the interaction patterns and approximate flow of communication in a given dialog. All turn-taking features were extracted using timings recorded for each utterance in the transcripts. Finally, we also incorporated ratio-based features to measure other aspects of the interaction patterns. We included the following ratio-based features:

- **Question Ratio:** The number of DAs tagged as *Request:Clarification* or *Question:General* from participants, normalized by all DA tags in the dialog.
- **Answer Ratio:** The number of DAs tagged as *Answer:General*, *Answer:Yes*, *Answer:No*, or *Acknowledgement* by an interviewer, normalized by all DA tags in the dialog.

These features were designed to capture various aspects of global interaction patterns that may have been missed by other feature groups, and have also shown promise in prior work on specific tasks (Nasreen et al., 2021; Khodabakhsh et al., 2015).

## 4 Data

### 4.1 Data Sources

We evaluated our DA tagging and AD detection models on two tasks in a subset of *DementiaBank*

| | Cookie | Fluency |
|---|---|---|
| **Gender** | M=45, F=52 | M=9, F=8 |
| **Age** | 90.29±35.01 | 66.47±8.41 |
| **Education** | 13.10 ±2.65 | 12.59±2.99 |
| **Onset Age** | 65.31±8.64 | 63.88±8.23 |

Table 1: Demographic information for both tasks. *Cookie* refers to the picture description task, and *Fluency* refers to the verbal fluency task. *Education* is in years. The *onset age* is the age of a participant when first diagnosed with AD.

known as the *Pitt corpus* (Becker et al., 1994). DementiaBank is a large database encompassing corpora pertaining to dementia submitted by numerous contributors around the globe. It includes corpora in multiple languages, spanning multiple cognitive tasks. The Pitt corpus is an English-language subset containing longitudinal dementia and control audiorecordings and associated transcripts for four language tasks, including picture description, verbal fluency, sentence construction, and story recall. We selected the picture description and verbal fluency tasks for our experiments.

The picture description task, known formally as the *Cookie Theft Picture Description Task* (Roth, 2011), is the most commonly studied task in research towards automated dementia detection, serving as the focus of two popular challenges in 2020 (Luz et al., 2020) and 2021 (Luz et al., 2021). It includes semi-structured interviews between an interviewer and a subject belonging to one of two groups (AD or non-AD). The subject is instructed to describe the contents of an eventful picture featuring, among other things, a child stealing a cookie. Previously, Farzana et al. (2020) annotated 100 transcripts from this dataset spanning 1616 utterances with 26 DA tags. The DA classes were adapted from the ISO Standard 24617-2 (Bunt, 2006) DA scheme, with the addition of 8 task-specific DAs.

The second task, designed to assess verbal fluency, features dialog between a participant and an interviewer. In the first segment of the interview, the participant is prompted to utter as many animal names as they can in one minute. In the second segment, they are instructed to utter as many words starting with *f* as they can within one minute.

For AD detection using the DA tags, we filtered the dataset such that each subject had one conversation.[1] We excluded three annotated conversations from Farzana et al. (2020)'s *Cookie-Theft DA* corpus, since two of the conversations belonged to a repeated participant (in different years) and the other's participant overlapped with one also present in the verbal fluency task. Altogether, our final dataset included 97 conversations (*non-AD*=46, *AD*=51) from the picture description task, and 17 (*non-AD*=2, *AD*=15) from the fluency task.[2] The annotations are available for the research community[3] for further followup work, and can be used after separately gaining access to DementiaBank.[4] Table 1 presents demographic statistics and Table 2 presents descriptive statistics for each task.

| | Cookie | Fluency |
|---|---|---|
| **# Conversations** | 97 | 17 |
| **Total Utterances** | 1569 | 760 |
| **Average Duration** | 672.43 | 147.29 |
| **Words/Minute** | 623.16 | 300.19 |

Table 2: Descriptive statistics for both tasks. Duration is in seconds, averaged across the number of conversations in the task.

### 4.2 Data Annotation

Although we were able to use the existing DA tags from *Cookie-Theft DA* directly, we manually annotated the 17 transcripts from the verbal fluency task with corresponding DAs. We followed the same guidelines established by Farzana et al. (2020), with minor task-specific adjustments. Specifically, we replaced the 8 task-specific DA tags corresponding to core topics in the cookie theft picture (denoted with labels *Answer:t1–Answer:t8* in *Cookie-Theft DA*) with two task-specific DA tags more closely aligned with the verbal fluency task; namely, *Answer:Topic1* and *Answer:Topic2*. *Answer:Topic1* is assigned to utterances in which participants refer to animal names, and *Answer:Topic2* is assigned to utterances in which participants say words beginning with the letter *f*. We distinguished these tags from one another to facilitate easy sepa-

[1] Since the Pitt corpus contains longitudinal data, some subjects have multiple entries for the same task, from initial and follow-up visits.

[2] We annotated 17 transcripts from verbal fluency task in *DementiaBank*, which is an imbalanced corpus with 2 and 239 transcripts from the *non-AD* and *AD* classes respectively, to avoid having a huge class imbalance in our resulting dataset.

[3] https://nlp.lab.uic.edu/resources/

[4] https://dementia.talkbank.org/

| DA | Label | Example | Ratio |
|---|---|---|---|
| QUESTION: GENERAL | *qg* | do you know other types? | <0.1 |
| QUESTION: REFLEXIVE | *qr* | a bird? | <0.1 |
| ANSWER: YES | *ay* | yeah that's fine | <0.1 |
| ANSWER: NO | *an* | I don't know | <0.1 |
| ANSWER: GENERAL | *ag* | gosh I can't think of it | <0.1 |
| INSTRUC-TION | *is* | words that begin with f | 0.2 |
| SUGGEST. | *sg* | just keep naming them | <0.1 |
| ACK. | *ak* | okay good | 0.1 |
| REQUEST: CLAR. | *rc* | did I say facts? | <0.1 |
| FEEDBACK: REFLEXIVE | *fr* | no that's not an animal | <0.1 |
| STALLING | *sl* | oh let's see | <0.1 |
| OTHER | *or* | &=laughs | <0.1 |
| ANSWER: TOPIC | *at* | uh dog, &hm oh a fence | 0.5 |

Table 3: DAs with non-zero frequency in our *Verbal Fluency DA* dataset, with examples. For DA tagging and AD detection, we reduce the task-specific DAs in both tasks to *Answer:Topic*. *Ratio* indicates the specified DA's frequency ratio for the *verbal fluency* task.

ration of tasks in later analyses.

Two graduate students annotated these transcripts adhering to the annotation guidelines published by Farzana et al. (2020), with new amendments added for the task-specific DA classes, after an initial training session with a practice transcript. They achieved strong inter-annotator agreement, as measured using Cohen's kappa (Cohen, 1960) with a score of $\kappa = 0.79$. The annotations were collected using the INCEpTION framework (Klie et al., 2018), a free, user-friendly, web-based annotation interface with built-in support for adjudication and assessment of inter-annotator agreement. Disagreements were forwarded to a third-party, expert adjudicator for final label selection. Table 3 presents example labeled utterances from our new *Verbal Fluency DA* corpus with a variety of DA tags.

## 5 Experiment

We conducted two core sets of experiments in this work. In the first set (§5.1), we evaluated the performance of our DA tagging model (described in §3.1) at correctly assigning labels from our annotation scheme to utterances in the picture description and verbal fluency tasks. In the second (§5.2), we measured the performance of features designed to capture meaningful interaction patterns using these DAs when leveraged in a dementia detection task.

### 5.1 DA Tagging

To evaluate the performance of our DA tagging model, we devised a series of experimental conditions featuring different components of interest in our study:

- **NO-CONTEXT:** The current utterance embedding. This was used as our baseline model.
- **n EMB.:** An utterance embedding history of length $n$ is passed to the DA prediction model. For example, when $n = 1$, the current utterance is passed to the model, and when $n = 2$, the previous utterance is used as context along with the current one.
- **n POS:** A POS embedding history of length $n$ is passed to the DA prediction model. For example, when $n = 1$, the POS tag sequence for the current utterance is passed to the model, and when $n = 2$, the POS tags for the previous utterance are used as context along with the current sequence.
- **n SPK.:** A speaker history of length $n$ is passed to the DA prediction model. For example, when $n = 1$, the current speaker tag is passed to the model, and when $n = 2$, the speaker tag for the previous utterance is used as context along with the current speaker tag.

Studying performance under these different conditions allowed us to develop a fuller understanding of the contributions of individual components. To implement our DA tagging model, we used a neural network backbone with the fine-tuned hyperparameters: *learning rate* = 0.001, *Adam* optimizer (Kingma and Ba, 2015), *batch size* = 32, *epoch*

| Feature | Details |
|---|---|
| **Unigram** | (*I_Instruction*), (*P_Request:Clarification*) |
| **Bigram** | (*P_Request:Clarification* + *I_Answer:General*), (*P_Stalling* + *I_Acknowledgment*) |
| **Trigram** | (*I_Instruction* + *P_Request:Clarification* + *I_Answer:General*) |
| **Ratio + Turn-Taking** | *Question Ratio*, *Answer Ratio*, *Average Turn Duration*, *Normalized Turn Switch*, *Total Duration*, *Average Words/Minute* |

Table 4: Interaction pattern features for AD detection. Durations are in milliseconds (ms).

| Features | Joint | Cookie | Fluency |
|---|---|---|---|
| 1 EMB. (NO CONTEXT) | 0.77 | 0.82 | 0.71 |
| 1 EMB. & 1 POS & 1 SPK. | 0.77 | 0.82 | 0.73 |
| 2 EMB. | 0.81 | 0.84 | 0.74 |
| 2 EMB. & 2 POS & 2 SPK. | 0.81 | 0.83 | 0.74 |
| 1 EMB. & 1 POS | 0.78 | 0.82 | 0.70 |
| 2 EMB. & 2 POS | 0.79 | 0.84 | 0.75 |
| 1 EMB. & 1 SPK. | 0.78 | 0.81 | 0.74 |
| **2 EMB. & 2 SPK.** | **0.82** | **0.85** | **0.75** |

Table 5: 10-fold cross-validation DA tagging results with micro-averaged $F_1$ scores on the *picture description* (Cookie), *verbal fluency* (Fluency) and *joint* tasks.

= 300, and early stopping criteria *min* $\delta = 0.0001$. We used a class-weighted categorical cross-entropy loss, since our class labels (for both the *picture description* and *verbal fluency* tasks) are imbalanced. Our utterance embeddings were computed using the *nli-bert-large* (Conneau et al., 2017) model with an embedding dimension of 1024 from the HuggingFace *sentence-transformers* library.

Finally, to capture our DA tagging model's ability to generalize, we also compared three versions of each condition. Specifically, we trained and evaluated the DA tagger on the *picture description* and *verbal fluency* tasks separately, and then we also trained and evaluated a *joint* model using data from both the tasks combined. This allowed us to empirically validate the feasibility of this model in several settings for later use in extracting AD detection features.

### 5.2 AD Detection

To evaluate the impact of our interaction features on classifying AD status in a task-agnostic setting, we performed experiments considering the following conditions:

- **ALL:** This condition utilizes all features included in Table 4 and described previously.
- **N-GRAM:** This condition includes all features in the rows corresponding to unigrams, bigrams, and trigrams in Table 4.
- **N-GRAM + TURN-TAKING:** This condition includes the union of all n-gram and interaction features listed in Table 4.
- **N-GRAM + RATIO:** This condition includes the union of all n-gram and ratio features listed in Table 4.

We implemented our AD detection model using a random forest classifier (*rfc*) with the following hyperparameters: *number of estimators*=100, *max depth*=10. We selected this model from among a pool of it and two other feature-based classification models (support vector models with polynomial and radial basis functions, respectively) based on preliminary performance validation experiments. Our choice of a feature-based classifier rather than more complex (and potentially higher-performing) neural network alternatives was driven by our need for easy interpretability, to analyze and compare features in task-agnostic settings.

## 6 Results

### 6.1 DA Tagging

We summarize the results of our DA prediction experiments in Table 5, comparing all conditions earlier described in §5.1. For the **n EMB.**, **n POS**, and **n SPK.** conditions, we use values of $n \in \{1, 2\}$. We limited our experiment to include only the immediate previous context ($n = 2$) since

| | Accuracy | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|
| | | AD | HC | AD | HC | AD | HC |
| **BASELINE** | 0.58 | 1.00 | 0.00 | 1.00 | 0.00 | 0.73 | 0.00 |
| **ALL** | **0.79** | 0.80 | **.77** | 0.85 | **0.71** | **0.82** | **0.74** |
| **N-GRAM** | 0.68 | 0.72 | 0.63 | 0.74 | 0.60 | 0.73 | 0.62 |
| **N-GRAM + TURN-TAKING** | 0.74 | 0.76 | 0.70 | 0.79 | 0.67 | 0.78 | 0.68 |
| **N-GRAM + RATIO** | 0.71 | 0.76 | 0.65 | 0.73 | 0.69 | 0.74 | 0.67 |

Table 6: Five-fold cross-validation results, for models jointly trained on the *picture description* and *verbal fluency* tasks using gold-annotated DA tags. The baseline model predicted the most frequent class for each instance.

that contributed the strongest performance boost in prior research (Nasreen et al., 2021; Farzana et al., 2020). Our baseline model (NO CONTEXT) yielded micro-averaged $F_1$ scores of $F_1$=0.77, $F_1$=0.82, and $F_1$=0.71 on the *joint*, *picture description*, and *verbal fluency* training settings, respectively. The performance of $F_1$=0.82 for the *picture description* setting exceeds that of the highest-performing benchmarking model reported by Farzana et al. (2020).

The results were further improved by adding contextual information from previous utterances (2 EMB.), achieving scores of $F_1$=0.81, $F_1$=0.84, and $F_1$=0.74 on the *joint*, *picture description*, and *verbal fluency* training settings, respectively. Adding the previous utterance's POS sequences and speaker tag (2 EMB. & 2 POS & 2 SPK.) did not offer noticeable advantages beyond this, with nearly equivalent performance. Overall, we observe the strongest performance when contextual embeddings and speaker tags are used *without* contextual part-of-speech sequences (2 EMB. & 2 SPK.), achieving scores of $F_1$=0.82, $F_1$=0.85, and $F_1$=0.75 on the *joint*, *picture description*, and *verbal fluency* training settings, respectively.

When comparing training settings, we observe that the *picture description* setting consistently achieves the highest performance, followed by the *joint* setting and finally the *verbal fluency* setting. This makes sense intuitively. The *verbal fluency* dataset was the smallest, and its size may have interfered with the DA prediction model's ability to derive meaningful information from the feature set. The *joint* dataset was the largest, but it may have struggled to effectively distinguish between class traits that manifested differently in different tasks. The *picture description* task is the most well-studied, and the only one for which benchmarking results were available (Farzana et al., 2020). We note that all of our models exceeded Farzana et al. (2020)'s strongest benchmark ($F_1$=0.77).

### 6.2 AD Detection

We summarize the results of our AD detection experiments in Table 6. For these results, we employed the *joint* corpus and used a five-fold cross-validation training and evaluation setting, comparing the different feature combinations outlined in §5.2. We report precision, recall, and $F_1$ for each class (AD and healthy control participants without AD, referred to as HC), as well as overall accuracy. We observe the highest performance under the ALL condition, with per-class $F_1$ scores of $F_1$=0.82 and $F_1$=0.74 for the AD and HC classes, respectively, and an overall accuracy of 0.79. This provides evidence of meaningful contributions from all interaction features when used in a task-agnostic AD detection setting.

All AD detection models exceeded the baseline condition (predicting the most frequent class, AD, in all cases). When combined with the DA tag unigrams, bigrams, and trigrams, the turn-taking features (accuracy=0.74) outperformed the ratio-based features (accuracy=0.71), although both added utility beyond the DA tag n-grams alone (accuracy=0.68). At a per-class level, performance for the AD class exceeded that of the control class; this was expected given that the dataset included a higher percentage of AD than HC participants.

### 6.3 Discussion

The results observed from the AD detection experiments clearly suggest that features based on content-free interaction patterns are helpful for dementia detection classifiers in task-agnostic set-

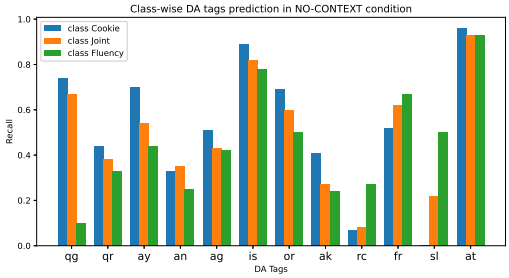


Figure 2: Comparison of class-wise recall of DA tags in the NO CONTEXT condition for all three tasks.

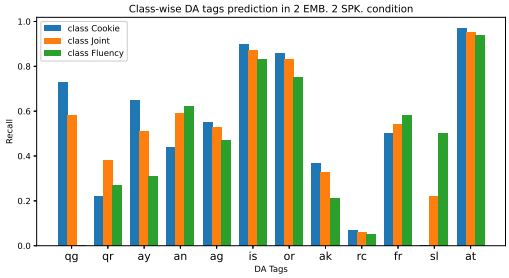


Figure 3: Comparison of class-wise recall of DA tags in the 2 EMB. 2 SPK. condition for all three tasks.

tings. Using only these features, our AD detection model was able to achieve performance comparable to that seen with content-driven, task-specific alternatives (Luz et al., 2020; Di Palo and Parde, 2019). This holds exciting implications for downstream diagnostic or assessment applications, which may be able to leverage these more general features rather than retraining models for new tasks.

To further understand the performance of our DA prediction model and examine the extent to which automated DA tags can support AD detection, we conducted additional error analyses. Specifically, we investigated model outcomes for different DA tags in the baseline (NO CONTEXT) and highest-performing conditions across the *picture description*, *verbal fluency*, and *joint* task settings. We illustrate the findings from these analyses in Figures 2, 3, and 4. Overall, we observed poor recall scores (Figure 2 and 3) for the *Request:Clarification* (*rc*) tag in both models across all tasks. This may be because *rc* utterances can be easily confused with *Question:Reflexive* (*qr*) or *Question:General* (*qg*) tags since they carry similar linguistic and syntactic characteristics (Farzana et al., 2020). Although these question types differ in their intent (*rc* conveys follow-up questions or lack of understanding of specific prior context, whereas *qr* is observed in

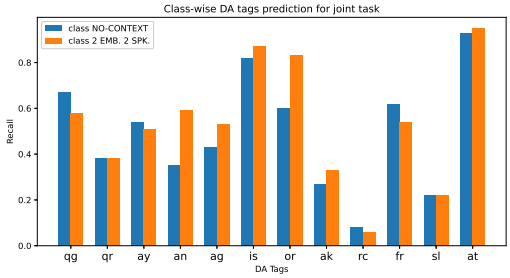


Figure 4: Comparison of class-wise recall of DA tags in NO-CONTEXT vs. 2 EMB. 2 SPK. for the *joint* task.

think-aloud scenarios during which subjects question themselves and *qg* is most commonly seen in out-of-context queries), they ultimately all seek information in some form.

When comparing the NO CONTEXT and highest-performing conditions across all tasks, we also observed that, surprisingly, adding prior context was not always beneficial. In the case of *rc* specifically, performance degrades when the previous speaker tag and utterance embedding are added, primarily in the *verbal fluency* task. The same pattern holds true for *qg* in the *verbal fluency* task, and *qr* in the *picture description* task.

Nonetheless, prior context boosts model performance (or has no negative impact) on a variety of DA classes across tasks. Figure 4 captures the effect of having speaker tags and utterance embeddings both for the current and previous utterance, and shows increases in recall for *Instruction* (*is*), *Other* (*or*), *Acknowledgement* (*ak*), *Answer:No* (*an*), and *Answer:General* (*ag*); we note that these dialog classes in general are associated with utterances that are strongly situated in context. For example, *Instructions* may differ in form depending on how they are received, and *Answer:General* and *Answer:No* are mostly uttered in the context of a question in the previous utterance. Utterances labeled as *Other* are mostly out-of-context statements or non-verbal expressions made by participants, and therefore the inclusion of speaker information is helpful in understanding these utterances.

## 7 Conclusion

In this paper we studied the extent to which automated analyses of interaction patterns can be leveraged for task-agnostic dementia detection. To do so, we adapted a DA annotation scheme for two different cognitive tasks. We then presented a context-aware DA tagging model that uses transfer learning

from pre-trained sentence embeddings to compute rich representations of utterances, paired with linguistic features (POS sequences) and speaker tags. The model achieved scores of $F_1$=0.75, $F_1$=0.85, and $F_1$=0.82 in a *verbal fluency*, *picture description*, and *joint* task, respectively. We find that although performance is low for some rare-class DAs, adding context information and speaker tags boosts performance in several cases.

To test the utility of interaction patterns as content-free features, we generate features based on these DA tags and other interaction characteristics. We use these to train a random forest classifier for task-agnostic AD detection and achieve strong performance on a joint dataset of *picture description* and *verbal fluency* dialogs. These interpretable interaction features in cognitive health screening tasks show promising performance in AD detection. In the future, we will extend this work to create a more balanced (across tasks and AD vs. non-AD classes) cognitive screening dataset, to further test the boundaries to which these results may generalize. These findings will allow us to outline a guiding principal for designing dialog agents for virtual interviewing in the cognitive health screening domain.

## 8 Acknowledgment

This work was supported in part by a startup grant from the University of Illinois Chicago. We thank Nitish Dewan for contributing to the data annotation process, and the anonymous reviewers for their helpful feedback.

## References

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.

Harry Bunt. 2006. Dimensions in dialogue act annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Flavio Di Palo and Natalie Parde. 2019. Enriching neural models with targeted features for dementia detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 302–308, Florence, Italy. Association for Computational Linguistics.

Christopher Elsey, Paul Drew, Danielle Jones, Daniel Blackburn, Sarah Wakefield, Kirsty Harkness, Annalena Venneri, and Markus Reuber. 2015. Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics. *Patient Education and Counseling*, 98(9):1071–1077.

Shahla Farzana, Ashwin Deshpande, and Natalie Parde. 2022. How you say it matters: Measuring the impact of verbal disfluency tags on automated dementia detection. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 37–48, Dublin, Ireland. Association for Computational Linguistics.

Shahla Farzana and Natalie Parde. 2020. Exploring MMSE Score Prediction Using Verbal and Non-Verbal Cues. In *Proc. Interspeech 2020*, pages 2207–2211.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Heidi Ehernberger Hamilton. 1994. *Conversations with an Alzheimer's Patient: An Interactional Sociolinguistic Study*. Cambridge University Press.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 928–939, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder Institute of Cognitive Science, Boulder, CO.

Ali Khodabakhsh, Fatih Yesil, Ekrem Guner, and Cenk Demiroglu. 2015. Evaluation of linguistic and prosodic features for detection of alzheimer's disease

in turkish conversational speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):9.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2020. Alzheimer's Dementia Recognition Through Spontaneous Speech: The ADReSS Challenge. In *Proc. Interspeech 2020*, pages 2172–2176.

Saturnino Luz, Fasih Haider, Sofia de la Fuente, Davida Fromm, and Brian MacWhinney. 2021. Detecting Cognitive Decline Using Speech Only: The ADReSSo Challenge. In *Proc. Interspeech 2021*, pages 3780–3784.

Jirí Martínek, Christophe Cerisara, Pavel Král, and Ladislav Lenc. 2021. Cross-lingual approaches for task-specific dialogue act recognition. In *AIAI*.

Bahman Mirheidari, Daniel Blackburn, Traci Walker, Markus Reuber, and Heidi Christensen. 2019. Dementia detection using automatic analysis of conversations. *Computer Speech & Language*, 53:65–79.

Shamila Nasreen, Julian Hough, and Matthew Purver. 2021. Rare-class dialogue act tagging for Alzheimer's disease diagnosis. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 290–300, Singapore and Online. Association for Computational Linguistics.

Shamila Nasreen, Matthew Purver, and Julian Hough. 2019. A corpus study on questions, responses and misunderstanding signals in conversations with alzheimer's patients. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

J. B. Orange, Rosemary B. Lubinski, and D. Jeffery Higginbotham. 1996. Conversational repair by individuals with dementia of the alzheimer's type. *Journal of Speech, Language, and Hearing Research*, 39(4):881–895.

Ulla Petti, Simon Baker, and Anna Korhonen. 2020. A systematic literature review of automatic Alzheimer's disease detection from speech and language. *Journal of the American Medical Informatics Association*, 27(11):1784–1797.

Charlene Pope and Boyd H. Davis. 2011. Finding a balance: The carolinas conversation collection. *Corpus Linguistics and Linguistic Theory*, 7(1):143–161.

P.A. Pérez-Toro, S.P. Bayerl, T. Arias-Vergara, J.C. Vásquez-Correa, P. Klumpp, M. Schuster, Elmar Nöth, J.R. Orozco-Arroyave, and K. Riedhammer. 2021. Influence of the Interviewer on the Automatic Assessment of Alzheimer's Disease in the Context of the ADReSSo Challenge. In *Proc. Interspeech 2021*, pages 3785–3789.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Carole Roth. 2011. Boston diagnostic aphasia examination. In Jeffrey S. Kreutzer, John DeLuca, and Bruce Caplan, editors, *Encyclopedia of Clinical Neuropsychology*, pages 428–430. Springer New York, New York, NY.

Utkarsh Sarawgi, Wazeer Zulfikar, Nouran Soliman, and Pattie Maes. 2020. Multimodal inductive transfer learning for detection of alzheimer's dementia and its severity. In *Proc. Interspeech 2020*, pages 2212–2216.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Ana Varela Suárez. 2018. The question-answer adjacency pair in dementia discourse. *International Journal of Applied Linguistics*, 28(1):86–101.

Bishan Yang, Jian-Tao Sun, Tengjiao Wang, and Zheng Chen. 2009. Effective multi-label active learning for text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, page 917–926, New York, NY, USA. Association for Computing Machinery.

Zining Zhu, Jekaterina Novikova, and Frank Rudzicz. 2019. Detecting cognitive impairments by agreeing on interpretations of linguistic features. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1431–1441, Minneapolis, Minnesota. Association for Computational Linguistics.