Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1167–1177
Marseille, 11–16 May 2020

Modeling Dialogue in Conversational Cognitive Health Screening Interviews

Shahla Farzana, Mina Valizadeh, Natalie Parde
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan St., Chicago, IL 60607
{sfarza3, mvaliz2, parde}@uic.edu

Abstract

Automating straightforward clinical tasks can reduce workload for healthcare professionals, increase accessibility for geographically-isolated patients, and alleviate some of the economic burdens associated with healthcare. A variety of preliminary screening procedures are potentially suitable for automation, and one such domain that has remained underexplored to date is that of structured clinical interviews. A task-specific dialogue agent is needed to automate the collection of conversational speech for further (either manual or automated) analysis, and to build such an agent, a dialogue manager must be trained to respond to patient utterances in a manner similar to a human interviewer. To facilitate the development of such an agent, we propose an annotation schema for assigning dialogue act labels to utterances in patient-interviewer conversations collected as part of a clinically-validated cognitive health screening task. We build a labeled corpus using the schema, and show that it is characterized by high inter-annotator agreement. We establish a benchmark dialogue act classification model for the corpus, thereby providing a proof of concept for the proposed annotation schema. The resulting dialogue act corpus is the first such corpus specifically designed to facilitate automated cognitive health screening, and lays the groundwork for future exploration in this area.

Keywords: dialogue act classification, dialogue modeling, cognitive health screening

1. Introduction

Recent advancements in artificial intelligence have opened new pathways for improving patient and clinician healthcare experiences, with technologies including but not limited to predictive disease modeling, ambient healthcare monitoring, and clinical record-keeping assistance. At the same time, shifting population demographics have ushered in new and pressing healthcare concerns. Dementia is one such increasingly critical concern as median population ages around the globe continue to rise (Tom et al., 2015). Although cures for dementia remain out of reach, researchers believe that early diagnosis can mitigate its effects (Prince et al., 2011). Diagnosis typically requires cognitive tests or in-person screening interviews, which can be costly, resource-intensive, and stressful for patients. A conversational agent capable of conducting screening interviews to elicit the information necessary to perform a preliminary assessment of cognitive health could pose an inexpensive, flexible, low-stress alternative that could simultaneously increase patient accessibility and reduce clinician workload.

Training such an agent to interpret natural language input and select suitable follow-up responses cannot be done without appropriate, clinically-relevant dialogue modeling data. Currently, there exists no spoken or text-based corpus containing utterances annotated with the dialogue intentions and associated characteristics necessary to facilitate cognitive health screening. In this work, we set out to fill that void. Our contributions are as follows:

1. We establish a dialogue act annotation schema for a popular, clinically-validated conversational cognitive health screening task.

2. Using that schema, we collect dialogue act annotations for an existing collection of transcribed conversations for that task. This data source is commonly used for automated dementia detection (Habash et al., 2012; Orimaye et al., 2014; Fraser et al., 2016; Yancheva and Rudzicz, 2016; Orimaye et al., 2018; Karlekar et al., 2018; Di Palo and Parde, 2019), and thus we anticipate that our additional layer of dialogue act annotations will be of broader interest to those working on automated dementia detection as well.

3. We demonstrate that the resulting corpus exhibits high inter-annotator agreement.

4. We establish a benchmark dialogue act classification model, validating and providing a proof of concept for the annotation schema.

Notably, our corpus is the first dialogue act corpus specifically designed to facilitate automated cognitive health screening. The rest of the paper is organized as follows. We summarize relevant dialogue act annotation and dementia diagnosis literature in Section 2. In Section 3, we detail our annotation schema and data collection process. We analyze the resulting dialogue act corpus in Section 4. In Section 5, we establish a benchmark dialogue act classification model. Finally, in Section 6, we summarize our findings and briefly describe our future plans.

2. Background

2.1. Dialogue Act Annotation

Dialogue act (DA) annotation is the process by which functionally- and contextually-appropriate labels are assigned to spans of dialogue, or *utterances*. The rules defining the set of DAs accepted for a given domain or task are referred to as the DA *annotation schema*. Over the years, many DA annotation schemata have been developed for conversational and task-based interactions. Early examples include TRAINS (Allen et al., 1995) and DAMSL

(Core and Allen, 1997) in the United States, Map Task in the United Kingdom (Anderson et al., 1991), and Verbmobil in Germany (Alexandersson et al., 1997).
Researchers used those early guidelines and others to construct a variety of spoken dialogue corpora. These corpora included the task-oriented Map Task Corpus (Anderson et al., 1991), the multimodal AMI Meeting Corpus (Mccowan et al., 2005), and the conversational SWITCHBOARD corpus (Godfrey et al., 1992). Although these corpora fueled the burgeoning tasks of automated dialogue act classification and subsequent dialogue management, an underlying weakness was their lack of consistency with one another. More recently, many researchers have coalesced upon using variations of the ISO Standard 24617-2 (Bunt, 2011), a portable, application-independent annotation schema that can adequately deal with typed, spoken, and multimodal dialogue. We adapt our schema from the ISO Standard 24617-2 to foster compatibility with recent (Fang et al., 2012; Bunt et al., 2016; Petukhova et al., 2014; Petukhova et al., 2018) and future corpora in other dialogue domains.

2.2. Health-Related Dialogue Act Corpora

Although many dialogue act corpora exist for both general conversation (e.g., SWITCHBOARD (Godfrey et al., 1992)) and specific tasks (e.g., tourist information (Young et al., 2010)), resources for work in the healthcare domain have thus far been scant. Gupta et al. (2018) recently released a corpus of 2858 SMS messages between patients and trained health coaches, annotated for specific goals and other DAs relevant to health behavior change therapy. Can et al. (2016) and Pérez-Rosas et al. (2016) focused on motivational interviewing dialogues between patients and therapists, coding counselor reflection types.[1]
Guntakandla and Nielsen (2018) released the only corpus thus far that has focused specifically on dialogues between trained interviewers and elderly patients. Similarly to Can et al. (2016) and Pérez-Rosas et al. (2016), they also examined reflection types. Their corpus includes reflection type labels (Complex Reflection, Simple Reflection, or No Reflection) for 1536 counselor utterances.
An underlying commonality of the corpora developed by Guntakandla and Nielsen (2018), Can et al. (2016), Pérez-Rosas et al. (2016), and Gupta et al. (2018) is that they all focus on psychological outcomes, either by directly promoting healthy behaviors (Gupta et al., 2018) or more indirectly promoting those behaviors by encouraging complex reflection during motivational interviews (Pérez-Rosas et al., 2016; Can et al., 2016; Guntakandla and Nielsen, 2018). In contrast, our focus is on facilitating cognitive assessment. Thus, we seek not to counsel or otherwise actively influence patients, but instead to encourage them to provide thorough narrative responses in the context of a natural conversation (Farzana and Parde, 2019).

2.3. Dementia Detection Corpora

Analysis of recorded or transcribed dialogue samples can provide valuable clues for early-stage dementia diagnosis.

[1]*Reflection*, a trademark characteristic of motivational interviewing, refers to the act of explicitly or implicitly mirroring content from the preceding utterance.

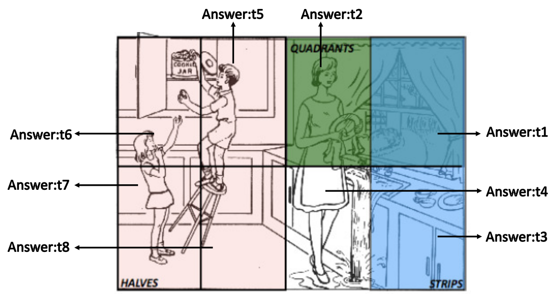

Figure 1: The image used for the *Cookie Theft Picture Description Task*. Annotated regions (Masrani, 2018) indicate different topic areas for our DA annotation schema.

Recent work in automated dementia detection has focused on such samples, harnessing linguistic, acoustic, and demographic features from text and/or audio data (Habash et al., 2012; Orimaye et al., 2014; Fraser et al., 2016; Yancheva and Rudzicz, 2016; Orimaye et al., 2018; Karlekar et al., 2018; Di Palo and Parde, 2019). Although our focus in this work is on dialogue modeling, our goal is specifically to model dialogue in the context of cognitive health assessment. Therefore, we collect dialogue act labels as an additional annotation layer for the most popular natural language dementia detection dataset, commonly known as *DementiaBank*. To provide requisite background, we summarize the publicly available datasets for dementia detection here.
DementiaBank (Becker et al., 1994) is the largest and most commonly-used dementia detection dataset. It consists of transcripts and recordings of English-, German-, Mandarin-, Spanish-, and Taiwanese-speaking participants with and without dementia completing several different language tasks. The task most frequently of interest to natural language processing researchers is the *Cookie Theft Picture Description Task*, a component of the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1972). In this task, English-speaking patients are asked to describe an image to an interviewer in a two-person conversational setting. We describe this data subset in more detail in Section 3.1, as it serves as the basis upon which our dialogue act corpus is built.
The *Western Aphasia Battery Dataset* (Risser and Spreen, 1985) contains writing samples that were elicited based on an image of a picnic. Subjects were asked to hand-write detailed descriptions of the scene, with the resulting descriptions filling a similar diagnostic role to that of the cookie theft picture descriptions. The *Cinderella Narrative Dataset* (Santos et al., 2017), used to detect mild cognitive impairment, contains speech samples elicited by asking participants to tell the "Cinderella" story after examining a corresponding picture book.
A common thread among these corpora is that the spoken and written samples were all collected manually by clinicians, during in-person visits. Not only is this costly in terms of clinicians' time; it is often inconvenient for patients, who must travel to testing sites at times dictated by clinicians' schedules. A conversational agent that can con-

duct these interviews automatically, on the patient's own time, at a location of their preference, could ameliorate these issues. Collecting a suitable DA corpus to train such an agent is a necessary first step toward its development, and the underlying motivation for our work here. Since DementiaBank is the largest existing source of cognitive health screening dialogues, in addition to being a popular resource among researchers for the downstream task of automated dementia detection, we select it as the basis upon which our dialogue act corpus is built.

2.4. Dialogue Act Modeling

Many studies conducted on dialogue seek to model its conversational structure by analyzing sequences of user intents. Intelligent systems designed to facilitate conversations autonomously need to replicate the same structure observed in natural human-human conversations, and to do so they typically manage input and select follow-up responses by classifying user utterances based on the dialogue act(s) that they realize (Shriberg et al., 1998; Prasad and Walker, 2002). Performing DA classification effectively enables the development of high-quality natural language dialogue systems (Higashinaka et al., 2014). A system's ability to accurately recognize different DAs often relies on a variety of information sources, including its own dialogue history.

DA recognition is known to be a complex problem, and many different approaches ranging from multi-class/multi-label classification to structured prediction have been applied to it (Stolcke et al., 2000; Yang et al., 2009). Rather than focusing on developing a state-of-the-art DA prediction model, our emphasis in this work is on the development of a DA corpus for cognitive health screening interviews. We validate the corpus here, in turn enabling us to shift our focus to the development of more complex DA recognition models in follow-up work. To validate our corpus, we train several well-known statistical and neural classification models that have been used previously for both DA prediction and other tasks; more details about our validation experiments can be found in Section 5. We refer readers to the wealth of additional studies on dialogue act modeling for more detailed studies focusing exclusively on DA classification (Venkataraman et al., 2003; Ang et al., 2005; Khanpour et al., 2016).

3. Methods

We built our dialogue act corpus on top of a popular existing dementia detection dataset, using a custom annotation schema designed to align well with schemata used in other dialogue domains. We describe the source corpus, annotation schema, and data collection process in the following subsections.

3.1. Source Corpus

We used the Pitt Corpus (Becker et al., 1994), a subset of DementiaBank, as the source for our dialogue act corpus. The Pitt Corpus contains verbal descriptions of an eventful image including, among other elements, a boy stealing a cookie (Figure 1). To elicit the descriptions, participants were shown the image and asked to describe what they saw. The interviewer coaxed participants to further elaborate their descriptions as needed. This process is known as the *Cookie Theft Picture Description Task*. The task is a clinically-validated assessment that was originally created for the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1972), and has been used in many clinical settings for dementia detection (Mendez and Ashla-mendez, 1991; Giles et al., 1996) and detection of other language impairments (Weintraub et al., 1990; Williams et al., 2010; Azambuja et al., 2012). The Pitt Corpus contains both audio recordings and transcripts for each participant who completed this task, as well as labels indicating their dementia status.

Group	Subjects	Transcripts	Avg. Words
AD	169	257	104.98 (SD=59.8)
MCI	19	43	111.09 (SD=55.8)
Control	99	242	113.56 (SD=58.5)

Table 1: Pitt Corpus statistics.

Overall, the Pitt Corpus includes 257 speech samples from participants diagnosed with probable or possible Alzheimer's disease or related dementia (AD), and 242 samples from healthy controls. It also contains a smaller number of speech samples from patients with mild cognitive impairment (MCI); these are patients with no official dementia diagnosis, but who received lower scores than healthy controls on a cognitive battery. Table 1 provides these statistics, along with average transcript length for each group.

The transcribed speech samples were separated into individual speech utterances (Interviewer=INV and Participant=PAR) following the CHAT-format annotation guidelines (MacWhinney, 2000). Although the transcripts also contain morphosyntactic information including part-of-speech tags, descriptions of tense, and repetition markers, we extracted only the speaker utterances (both INV and PAR). We did not remove or edit any CHAT transcription entities, such as indicators of filled pauses (e.g., "ah," "um"), repairs (e.g., "in the in the kitchen"), or non-standard word forms (e.g., "gonna"). Table 2 shows a sample transcript from the corpus, annotated using our dialogue act schema (described in the following subsection).

3.2. Annotation Schema

Our dialogue act annotation schema is based on the ISO Standard 24617-2 (Bunt, 2011), which inherits nine dimensions from the DIT^{++} scheme (Bunt, 2006). These dimensions, each of which cover different communicative activities, include: (1) *Task* DAs that move the communication forward, motivating the dialogue; (2-3) *Auto-* and *Allocation-Feedback* DAs, which provide or elicit information about the processing of previous utterances by the current speaker or addressee, respectively; (4) *Turn Management* DAs used to obtain or release the right to speak; (5) *Time Management* DAs that manage the use of time in the interaction; (6) *Discourse Structuring* DAs that deal with topic management and opening and closing sub-dialogues;

Speaker	DA	Utterance
INV	Instruction	this is the picture
INV	Instruction	just tell me what's happening in the picture
PAR	Answer: t5, Answer: t6	he's trying to steal cookie
PAR	Answer: t6	she's uh the little girl is uh saying shh
PAR	Answer: t2	uh the mother don't hear
PAR	Request: Clarification	did I tell say the sink was running over ?
INV	Answer: Yes	okay mhm you did
INV	Acknowledg.	okay that's fine

Table 2: Transcript fragment from the Pitt Corpus.

(7-8) *Own* and *Partner Communication Management* DAs, by which speakers can edit their contributions or the contributions of other speakers, respectively; and (9) *Social Obligations Management* DAs for dealing with social conventions (e.g., greetings, introductions, apologies, or expressions of gratitude). We adapted these dimensions to include custom roles necessary to the *Cookie Theft Picture Description Task*, to reflect the types of interactions typical of the task and to minimize complexity for our annotators; however, our adapted schema can be easily mapped to the original ISO Standard 24617-2.

In the *Task* dimension, we added labels ({*Answer:t1*-*Answer:t8*}) corresponding to the regions annotated in the image in Figure 1. We introduced these DA labels so that we could capture a finer-grained understanding of the comprehensiveness and coverage of a participant's description. These image regions have previously been validated as representative of the image's central *information units*, or themes (Masrani, 2018).

We simplified the DA hierarchy by removing the lowest-level distinctions for many dimensions, which are either difficult for novice annotators to judge (e.g., sub-types of *Time Management*), or can be recovered from other properties of the data. We also only included DAs from the following dimensions: *Task*, *Feedback*, *Time Management*, *Own/Partner Communication Management*, *Social Obligations Management*, and *Other* (a default for DAs not fitting into any other dimension). The dimensions are in principle often already independent of one another, and we explicitly instructed the annotators to assign only the most relevant DA label to each utterance, with the exception that multiple labels from {*Answer:t1*, ..., *Answer:t8*} could be assigned if necessary.[2] Our full adapted DA annotation schema is shown in Tables 3 and 4.

[2]For example, the utterance "the mother is wiping dishes and the water is running on the floor" could be labeled as *Answer:t2* or *Answer:t4*, but more accurately as both.

Task	
Question: General	Speaker wants information from addressee, and does not signal non-understanding.
Question: Reflexive	Speaker asks questions to him/herself, not to others.
Answer: Yes	Affirmative answer.
Answer: No	Negative answer.
Answer: General	Speaker provides complete or partial information in response to a question/instruction in a previous utterance.
Ans.: t1–t8	Illustrated in Figure 1.
Instruction	Speaker gives directions to do something, or makes a statement to elicit information from the addressee.
Suggestion	Speaker offers addressee an idea/plan for consideration.
Request	Speaker asks addressee to perform an action.
Offer	Speaker expresses readiness to do or give something to the addressee if desired.
Feedback	
Acknowl.	Speaker expresses understanding of the addressee.
Request: Clarification	Speaker asks a clarifying question regarding any previous context of the conversation, and expects a response from the addressee.
Feedback: Reflexive	Speaker answers his/her own questions or responds to his/her own statements.

Table 3: Dialogue act annotation schema (part one).

3.3. Data Collection

Two graduate students annotated 100 transcripts from the Pitt Corpus. The annotators viewed the entire transcripts and were asked to assign a DA label from the schema in Tables 3 and 4 to each segmented utterance. They received instruction on DA annotation as well as detailed guidelines and examples for the permitted labels. The annotations were collected using the WebAnno framework (Yimam et al., 2013), a free, user-friendly, web-based annotation interface. Annotators were told to choose the label corresponding to an utterance's main function, and were provided with an illustrated guide to the labels in {*Answer:t1*, ..., *An-*

Time Management	
Stalling	Speaker moderates the time needed to continue the dialogue directly or indirectly.
Own/Partner Communication Management	
Correction	Speaker corrects information from a previous utterance.
Social Obligation Management (SOM)	
Farewell	Speaker explicitly seeks to end a conversation. *Farewell* should not be confused with *Acknowledgement* — many transcripts may end without a *Farewell*.
Apology	Speaker desires to convey regret.
Greeting	Speaker explicitly seeks to begin a conversation. Many transcripts may begin without a *Greeting*.
Other	
Other	Default tag for otherwise non-classifiable utterances.

Table 4: Dialogue act annotation schema (part two).

DA	Speaker PAR	Speaker INV	Frequency
Answer:t1-Answer:t8	904	1	56%
Acknowledgement	34	156	11.8%
Instruction	0	128	7.9%
Answer:General	81	17	6.1%
Question:General	17	61	4.8%
Stalling	61	3	4.0%
Request:Clarification	49	11	3.7%
Answer:Yes	7	17	1.5%
Farewell	0	16	1.0%
Feedback:Reflexive	14	0	0.9%
Other	12	1	0.8%
Answer:No	5	4	0.6%
Question:Reflexive	9	0	0.6%
Correction	5	0	0.3%
Apology	2	1	0.2%
Grand Total (Count)	1616		

Table 5: DA frequencies, from highest to lowest.

swer:t8}. They were also told that those labels should take precedence over seemingly equally-applicable alternate labels (e.g., *Acknowledgement*).
Disagreements (including overlapping but non-identical sets of {*Answer:t1*, ..., *Answer:t8*} labels) were forwarded to a third-party, native English-speaking adjudicator. The adjudicator considered both annotations and selected the final label (or optionally, labels, if all were in {*Answer:t1*, ..., *Answer:t8*}) for the utterance.

4. Corpus Analysis

In total, the 100 annotated transcripts comprised 1616 distinct utterances. We assessed inter-annotator agreement using a modified version of Cohen's kappa (Cohen, 1960). More specifically, in cases when we allowed multiple labels for an utterance (e.g., with {*Answer:t1*, ..., *Answer:t8*}), we considered two annotators to agree if they had indicated at least one common label. We refer to this updated kappa statistic as *relaxed kappa* (κ_r), and present the updated formula below, letting x_m be a binary variable indicating agreement or disagreement regarding utterance m, S^i_m be the set of labels provided by annotator i for utterance m, N be the number of utterances, and n^i_k be the number of times annotator i predicted label k.

$$x_m = \begin{cases} 1, & \text{if } |S^1_m \cap S^2_m| > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$P_o = \frac{1}{N}\sum_m x_m \quad (2)$$

$$P_e = \frac{1}{N^2}\sum_k n^1_k n^2_k \quad (3)$$

$$\kappa_r = \frac{P_o - P_e}{1 - P_e} \quad (4)$$

Across all utterances, κ_r=0.75. This is in line with values reported for other recent DA corpora utilizing labels of similar granularity (e.g., Shirai and Fukuoka (2018)), and validates the feasibility of the annotation schema. We present all non-zero DA type frequencies in the corpus in Table 5, organized into both participant (PAR) and interviewer (INV) utterances. In Table 6, we also provide a breakdown of the ten most frequently co-occurring topic-related labels ({*Answer:t1*, ..., *Answer:t8*}). This was computed to facilitate analysis of which topics participants typically group together when describing the image; we observe that many frequent co-occurrences are spatially related.

4.1. Examples of Inter-Annotator Disagreement

In analyzing the collected data, we observe that certain words or phrases are generally more ambiguous and prone to causing confusion and disagreement than others. For example, a common source of disagreement among our annotators concerned phrases such as "alright thanks" or "okay that's fine," both of which could conceivably signal either *Acknowledgement* or *Farewell* at the end of the conversation. More generally, we found that label disagreement often occurred when words or phrases could be interpreted either way in a given context; often to more decisively disambiguate the speaker's true intent, prosodic or visual cues would be needed.
We provide an example of one such disagreement that occurred during data collection in Table 7. Specifically, the utterance *u10* of transcript *t057* can be interpreted either

DA	Count	Frequency
Answer:t5, Answer:t6	90	9.97%
Answer:t5, Answer:t8	59	6.53%
Answer:t2, Answer:t4	45	4.98%
Answer:t5, Answer:t6, Answer:t8	22	2.44%
Answer:t1, Answer:t2	18	1.99%
Answer:t3, Answer:t4	16	1.77%
Answer:t2, Answer:t5, Answer:t6	12	1.33%
Answer:t6, Answer:t8	11	1.22%
Answer:t6, Answer:t7	7	0.78%
Answer:t2, Answer:t3	6	0.66%

Table 6: The 10 most frequent co-occurring labels.

T. ID	U. ID	Speaker	Utterance
t057	u09	PAR	and the cookie jar's looking full.
t057	u10	INV	okay.
t057	u11	PAR	that's it.
t057	u12	INV	alright thanks.
t257	u00	INV	now I want you to tell me everything you see happening there.
t257	u01	INV	everything that you see going on in that picture.
t257	u02	PAR	&uh inside the room or every place ?

Table 7: Example disagreements, with columns indicating the transcript ID, utterance ID, speaker label, and utterance text.

as an *Acknowledgment* or as an expression of saying goodbye (*Farewell*), indicating the end of the conversation in response to the previous utterance of the participant. Transcript *t257* highlights another common source of disagreement in the corpus; namely, the task of disambiguating question types (*Question:General*, *Question:Reflexive*, or *Request:Clarification*). Correctly discerning the speaker's interrogative intent often depends on the context from the previous utterances. One can see how the utterance *u02* could conceivably be misunderstood as *Question:General* in isolation, whereas if it is placed in context, a more appropriate label becomes evident (*Request:Clarification*).

5. Dialogue Act Classification

To validate the utility of our corpus, we used it to train and evaluate a dialogue act classification model. In addition to demonstrating that the general and domain-specific speaker intents can be successfully modeled, this establishes a performance benchmark upon which we hope to improve in follow-up work. In this section we describe the features extracted for this benchmark, as well as the classification models considered in our experiments. We provide results showing which classification model exhibited the highest performance, and compare the performance of all experimental models with a baseline model that predicted the most frequent label (*Answer:t6*). Comparing with this latter baseline allowed us to validate that our model performed at a level clearly distinguishable from chance.

5.1. Features

For each utterance, we extract a vector of continuous (numeric) and categorical (one-hot encoded representations) features. These features can be subdivided into three categories: (1) **target utterance features**; (2) **context features**; and (3) **global dialogue features**. All features are derived from the interview transcripts and represent aspects of the dialogue in which each utterance occurs. Each feature category is described in further detail below.

5.1.1. Target Utterance Features

We extracted n-gram features for $n \in \{1, 2, 3\}$ from the entire training corpus, comprising all training utterances. We retained only those n-grams that appeared at least five times across the training data, and constructed a sparse feature vector for each utterance containing one dimension for each n-gram. Feature values were filled using TF-IDF counts for a given utterance, and each vector was L2-normalized with unit modulus.

5.1.2. Context Features

We introduced context features to model dependencies among consecutive utterances in a natural conversation. Specifically, we added the gold-standard DA labels[3] of the immediate previous utterance as a 26-dimensional one-hot encoded feature vector for the current utterance. This served as a simple way to address a key shortcoming of the standard multi-class classification models examined here (see Section 5.2.2); namely, that they are not naturally equipped to handle sequential information.

5.1.3. Global Dialogue Features

Finally, we incorporated global dialogue features as a mechanism for capturing speaker information rather than utterance-specific characteristics. The goal in including these features was to facilitate the classifier's ability to model turn-taking behavior, drawing upon our observations that certain dialogue acts (e.g., *Answer:t1-Answer:t8*) tend to be uttered by the participant in the interview, whereas others (e.g., *Instruction*) tend to be uttered by the interviewer. We refer the reader to Table 5 for an in-depth breakdown of participant and interviewer utterance distributions.

[3]We note that using the gold-standard labels reflects a hypothetical downstream case with perfect DA classification accuracy. Since this scenario may be unrealistic, we also include results without these features in Table 8 for comparison.

5.2. Experiments

5.2.1. Data Preprocessing

As previously noted, our corpus contains 1616 utterances across 100 interview transcripts, labeled using the 26 DAs described in Tables 3 and 4. All speakers in all transcripts are anonymous to protect user privacy; each transcript and each speaker within a given transcript were instead linked to a unique ID (transcripts are represented using an interview number, and speakers are identified using participant numbers). Participant demographic data (e.g., age, gender, and interview date) can be extracted from accompanying metadata files using the participant and transcript IDs. Sixty-four unique participants are represented among the 100 interview transcripts included in our corpus.

Our classification objective was to predict DA labels that matched the gold standard values (i.e., labels provided by trained annotators, or adjudicated labels in the event that the annotators disagreed). We partitioned the full dataset into 10 folds, with each fold containing 10 transcripts. We kept all transcripts belonging to the same participant in the same fold to ensure no unintentional biases or performance boosts in the classification results.

5.2.2. Classification Models

We experimented with multiple supervised machine learning methods to establish a strong performance benchmark for our dialogue act classification task. As exemplified earlier in Table 2, certain utterances in our corpus allow multiple labels (specifically, those containing topic-related information, i.e., {*Answer:t1-Answer:t8*}). This makes our classification task a multi-class, multi-label problem. In multi-label text classification, texts are automatically classified into a subset of one or more predefined classes, rather than strictly requiring a single class label. More formally, we can denote training examples (individual utterances) as $\{u_1, ..., u_n\}$ and the k classes (dialogue act labels) as $\{c_1, ..., c_k\}$. We can then represent the label set for utterance u_i as a binary vector $Y_i = [y_1, ..., y_k], y_j \in \{0,1\}$, where $y_j = 1$ if u_i belongs to class j or otherwise $y_j = 0$. We experiment with the following standard classification models for our multi-label dialogue act classification task.

- **Support Vector Classifier (SVC):** SVC has demonstrated significant success on text classification tasks (Joachims, 2002; Yang, 2001); hence, we include it in our experiments. Multi-label SVC adopts the one-versus-all approach, treating the DA classification task as multiple binary classification problems, where the utterances from the target class are given positive labels (i.e. $y = 1$), and the rest of the utterances are given negative labels (i.e., $y = 0$). Features were standardized by removing the mean and scaling to unit variance. Following prior work, we applied a linear kernel (Joachims, 1998) and kept the penalty parameter C at a default value of 1.0.

- **Decision Tree (DT):** Decision Trees hierarchically organize features based on information gain, and have proven to be successful on prior DA classification tasks (Moldovan et al., 2011; Stolcke et al., 2000); thus, we include it as an additional baseline classification model in our experiments.

- **Logistic Regression (LR):** Logistic Regression models have proved useful on numerous text classification tasks (Sankoff and Labov, 1979; Schütze et al., 1995; Berger et al., 1996; Ratnaparkhi, 1996; Kehler, 1997; Nigam et al., 1999). For this reason, we consider it as an additional baseline classification model.

For all standard classification models, we apply a one-versus-all wrapper to predict multiple labels, selecting any and all labels surpassing a prediction probability threshold of 0.5 and adhering to several post-processing constraints, outlined below. We implemented each classifier using sckikit-learn v0.21.3,[4] with the default hyperparameter settings for each. We used the same feature set (target utterance features, context features, and global dialogue features, described in Section 5.1) for all three standard classification models.

Our focus in establishing a benchmark was on standard classification algorithms due to the size of our corpus. However, we also conducted preliminary experiments to assess how neural networks perform compared to these other classifiers for our dataset. One potential advantage of neural models over standard classifiers is that their response function is continuous (smooth) at the decision boundaries, allowing them to avoid hard decisions and complete fragmentation of data associated with decision questions. For our experiments here, we used a simple feedforward neural network with a single hidden layer comprised of 128 units, and a sigmoid activation at the output layer. For label prediction, we then set our classification threshold at 0.5 for each class as usual. We used a binary cross-entropy loss function when optimizing the model to ensure that output nodes were penalized independently.

After predicting an initial set of labels based on probability thresholds, we passed the labels through a post-processing step to ensure that each of the previously-described models adhered to the same restrictions as our human annotators. Specifically, in this step we applied the following constraints:

- For utterances for which no predicted probability for any DA label exceeded the classification threshold, the model defaulted to selecting the DA label with the highest class probability. This was done to eliminate the possibility of an utterance having no predicted label, since leaving an utterance unlabeled was a disallowed behavior for human annotators.

- For utterances for which the predicted probability for multiple DA labels exceeded the classification threshold:[5]
  - If some were in {*Answer:t1*, ..., *Answer:t8*} and some were not, the model retained only those in {*Answer:t1*, ..., *Answer:t8*}.

[4] `https://scikit-learn.org/stable/`

[5] Human annotators were only allowed to select multiple labels if all of the selected labels were in {*Answer:t1*, ..., *Answer:t8*}.

– If none were in {*Answer:t1*, ..., *Answer:t8*}, the model retained the single label with the highest class probability.

5.3. Results

We evaluated performance for the most-frequent-label baseline (BASE), SVC, DT, LR, and neural network (NN) models using accuracy, Jaccard index, micro-averaged precision, micro-averaged recall, and micro-averaged F_1 score,[6] and present the results in Table 8. We calculated our performance metrics using 10-fold cross-validation across the entire dataset. As previously mentioned, we run experiments both with and without context features; cases without those features are denoted as -CONTEXT. Since some utterances contain multiple gold standard labels (e.g., *u01*={*Answer:t5*, *Answer:t6*}) and our accuracy metric requires an exact match in label(s) to count an utterance as a true positive, accuracy for the baseline model is closer to 9% despite the most frequent label making an appearance in the label set for 15% of utterances. We address this by measuring label overlap, and therefore partial matches (e.g., predictions of *Answer:t6* when the gold standard labels are {*Answer:t5, Answer:t6*}), using Jaccard index. Although rarely used to evaluate dialogue act classification models, Jaccard index is commonly used in other multi-label classification settings, and thus we include it alongside more standard single-class DA classification metrics (e.g., accuracy, precision, recall, and F_1) here. We compute Jaccard index across the full test set according to the equation below, where T is the set of one or more true labels for an utterance and P is the set of one or more predicted labels for the same utterance.

$$J(T,P) = \frac{1}{N}\sum_{k=1}^{N}\frac{T_k \bigcap P_k}{T_k \bigcup P_k} \tag{5}$$

Interestingly, we observe that the inclusion of context features results in only minor performance improvements across all models; thus, it appears that inter-utterance dependencies play a much smaller role than the text content itself and global characteristics when modeling dialogue in this domain. Despite our corpus being smaller than those typically used to train neural models, we find that NN outperforms the standard classification models across all metrics, establishing a strong performance benchmark for this dataset (Accuracy=68.64, Jaccard Index=75.54, Precision=80.59, Recall=75.06, and F_1=77.70). The three standard classification models exhibited similar performance to one another, with logistic regression slightly outperforming the others. All three easily outperformed the most frequent class baseline.

Model	Acc.	Jaccard	Precision	Recall	F_1
BASE	8.76	13.23	15.16	15.16	15.16
SVC	58.39	67.56	72.56	67.57	69.97
SVC-CONTEXT	58.15	66.57	71.79	66.76	69.16
DT	57.65	66.07	74.16	70.30	72.13
DT-CONTEXT	58.68	67.17	69.63	71.53	70.56
LR	65.30	72.77	79.97	70.32	74.79
LR-CONTEXT	65.24	72.09	79.24	70.06	74.31
NN	**68.64**	**75.54**	**80.59**	**75.06**	**77.70**
NN-CONTEXT	67.78	74.37	79.37	74.05	76.59

Table 8: Cross-validation results (%), where Precision, Recall and F_1 are micro-averaged. Note that in a multi-class setting where a single label is output for each class (i.e., BASE), micro-averaged precision and recall are expected to be the same (the sums of all false negatives and false positives, across all classes, will be equivalent to one another.)

6. Conclusions and Future Work

In this work, we build the first dialogue act corpus for cognitive health screening interviews. We design an annotation schema with 26 DA types corresponding to task-specific, goal-oriented, and general conversational cues, and collect labels adhering to this schema for 100 *Cookie Theft Picture Description* interviews between clinicians and elderly patients. The resulting corpus contains 1616 labeled utterances; we compute Cohen's kappa to assess inter-annotator agreeement and find that we achieve substantial agreement between two graduate student annotators (κ_r=0.75).

In analyzing the corpus, we find that the most common DA labels are task-specific ({*Answer:t1-Answer:t8*}), followed by a mixture of goal-oriented (*Instruction*) and traditional conversational (*Acknowledgement*) cues. This verifies our earlier observations that conversational cognitive health screening is characterized by a unique set of dialogue roles not adequately captured by existing schemata. As a proof of concept and to validate the feasibility of the corpus, we use the collected data to train a variety of statistical and neural classification models for DA classification, finding that all outperform a naïve baseline by a clear margin. Furthermore, we find that the top-performing model achieves high overall performance (Accuracy=68.64, Jaccard Index=75.54, Precision=80.59, Recall=75.06, and F_1=77.70), establishing a strong benchmark for this dataset. In the future, we plan to leverage this corpus to develop a conversational agent capable of facilitating cognitive health screening interviews, with the eventual goal of promoting greater healthcare accessibility for patients and reducing clinician burden. We publicly release our corpus[7] to the research community to stimulate additional interest in this field and foster further follow-up work from others.

[6]Due to the nature of our classification problem (multi-class, multi-label classification with many classes and a high class imbalance), we report micro-averaged metrics rather than individual scores for each class.

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

7. Bibliographical References

Alexandersson, J., Reithinger, N., and Maier, E. (1997). Insights into the dialogue processing of verbmobil. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, ANLC '97, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Allen, J. F., Schubert, L. K., Ferguson, G., Heeman, P., Hwang, C. H., Kato, T., Light, M., Martin, N., Miller, B., Poesio, M., and Traum, D. R. (1995). The trains project: a case study in building a conversational planning agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 7(1):7–48.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The hcrc map task corpus. *Language and Speech*, 34(4):351–366.

Ang, J., Yang Liu, and Shriberg, E. (2005). Automatic dialog act segmentation and classification in multiparty meetings. In *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I/1061–I/1064 Vol. 1.

Azambuja, M. J., Radanovic, M., Haddad, M. S., Adda, C. C., Barbosa, E. R., and Mansur, L. L. (2012). Language impairment in huntington's disease. *Arquivos de Neuro-psiquiatria*, 70(6):410–415.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of alzheimer's disease: Description of study cohort and accuracy of diagnosis. *Archives of Neurology*.

Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Bunt, H., Petukhova, V., Malchanau, A., Wijnhoven, K., and Fang, A. (2016). The dialogbank. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

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

Bunt, H. (2011). Multifunctionality in dialogue. *Comput. Speech Lang.*, 25(2):222–245, April.

Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., and Narayanan, S. S. (2016). "It sounds like...": A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of Counseling Psychology*, 63(3):343–350.

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

Core, M. G. and Allen, J. F. (1997). Coding dialogs with the damsl annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*.

Di Palo, F. and Parde, N. (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, July. Association for Computational Linguistics.

Fang, A., Cao, J., Bunt, H., and Liu, X. (2012). The annotation of the switchboard corpus with the new iso standard for dialogue act analysis. In *Workshop on Interoperable Semantic Annotation*, page 13.

Farzana, S. and Parde, N. (2019). Virtual-interviewer: A conversational agent designed to facilitate cognitive health screening in older adults. In *Proceedings of the IEEE-EMBS International Conference on Biomedical and Health Informatics: Extended Abstracts (BHI 2019)*, Chicago, Illinois, May.

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify alzheimer's disease in narrative speech. *Journal of Alzheimer's Disease*, 49 2:407–22.

Giles, E., Patterson, K., and Hodges, J. R. (1996). Performance on the boston cookie theft picture description task in patients with early dementia of the alzheimer's type: Missing information. *Aphasiology*, 10(4):395–408.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: telephone speech corpus for research and development. In *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1, March.

Goodglass, H. and Kaplan, E. (1972). *The assessment of aphasia and related disorders*. Lea & Febiger.

Guntakandla, N. and Nielsen, R. (2018). Annotating Reflections for Health Behavior Change Therapy. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Gupta, I., Di Eugenio, B., Ziebart, B., Liu, B., Gerber, B., Sharp, L., Davis, R., and Baiju, A. (2018). Towards building a virtual assistant health coach. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 419–421, June.

Habash, A., Guinn, C., Kline, D., and Patterson, L. (2012). *Language Analysis of Speakers with Dementia of the Alzheimer's Type*. Ph.D. thesis, 01.

Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T., and Matsuo, Y. (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, August. Dublin City University and Association for Computational Linguistics.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In Claire Nédellec et al., editors, *Machine Learning:*

ECML-98, pages 137–142, Berlin, Heidelberg. Springer Berlin Heidelberg.

Joachims, T. (2002). *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA.

Karlekar, S., Niu, T., and Bansal, M. (2018). Detecting linguistic characteristics of Alzheimer's dementia by interpreting neural models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 701–707, New Orleans, Louisiana, June. Association for Computational Linguistics.

Kehler, A. (1997). Probabilistic coreference in information extraction. In *Second Conference on Empirical Methods in Natural Language Processing*.

Khanpour, H., Guntakandla, N., and Nielsen, R. (2016). Dialogue act classification in domain-independent conversations using a deep recurrent neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2012–2021, Osaka, Japan, December. The COLING 2016 Organizing Committee.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, volume 1. Psychology Press.

Masrani, V. (2018). *Detecting dementia from written and spoken language*. Ph.D. thesis, University of British Columbia.

Mccowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska Masson, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research*.

Mendez, M. F. and Ashla-mendez, M. (1991). Differences between multi-infarct dementia and alzheimer's disease on unstructured neuropsychological tasks. *Journal of Clinical and Experimental Neuropsychology*, 13(6):923–932. PMID: 1779031.

Moldovan, C., Rus, V., and Graesser, A. (2011). Automated speech act classification for online chat. In *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference*, pages 23–29.

Nigam, K., Lafferty, J., and McCallum, A. (1999). Using maximum entropy for text classification. In *Proceedings of the IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

Orimaye, S. O., Wong, J. S.-M., and Golden, K. J. (2014). Learning predictive linguistic features for Alzheimer's disease and related dementias using verbal utterances. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 78–87, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Orimaye, S. O., Wong, J. S.-M., and Wong, C. P. (2018). Deep language space neural network for classifying mild cognitive impairment and alzheimer-type dementia. *PloS one*, 13(11):e0205636.

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. (2016). Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, San Diego, CA, USA, June. Association for Computational Linguistics.

Petukhova, V., Gropp, M., Klakow, D., Eigner, G., Topf, M., Srb, S., Motlicek, P., Potard, B., Dines, J., Deroo, O., Egeler, R., Meinz, U., Liersch, S., and Schmidt, A. (2014). The dbox corpus collection of spoken human-human and human-machine dialogues. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Petukhova, V., Malchanau, A., Oualil, Y., Klakow, D., Luz, S., Haider, F., Campbell, N., Koryzis, D., Spiliotopoulos, D., Albert, P., Linz, N., and Alexandersson, J. (2018). The Metalogue Debate Trainee Corpus: Data Collection and Annotations. In Nicoletta Calzolari (Conference chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018. European Language Resources Association (ELRA).

Prasad, R. and Walker, M. (2002). Training a dialogue act tagger for human-human and human-computer travel dialogues. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 162–173, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Prince, M., Bryce, R., Ferri, C., and Alzheimer's Disease International. (2011). World alzheimer report 2011: The benefits of early diagnosis and iserrntervention.

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*.

Risser, A. H. and Spreen, O. (1985). The western aphasia battery. *Journal of Clinical and Experimental Neuropsychology*, 7(4):463–470.

Sankoff, D. and Labov, W. (1979). On the uses of variable rules. *Language in society*, 8(2-3):189–222.

Santos, L., Corrêa Júnior, E. A., Oliveira Jr, O., Amancio, D., Mansur, L., and Aluísio, S. (2017). Enriching complex networks with word embeddings for detecting mild cognitive impairment from speech transcripts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1284–1296, Vancouver, Canada, July. Association for Computational Linguistics.

Schütze, H., Hull, D. A., and Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '95, pages 229–237, New York, NY, USA. ACM.

Shirai, K. and Fukuoka, T. (2018). JAIST annotated corpus of free conversation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Shriberg, E., Stolcke, A., Jurafsky, D., Coccaro, N., Meteer, M., Bates, R., Taylor, P., Ries, K., Martin, R., and van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41(3-4):443–492. PMID: 10746366.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373, September.

Tom, S. E., Hubbard, R. A., Crane, P. K., Haneuse, S. J., Bowen, J., McCormick, W. C., McCurry, S., and Larson, E. B. (2015). Characterization of dementia and alzheimer's disease in an older population: Updated incidence and life expectancy with and without dementia. *American Journal of Public Health*, 105(2):408–413. PMID: 25033130.

Venkataraman, A., Ferrer, L., Stolcke, A., and Shriberg, E. (2003). Training a prosody-based dialog act tagger from unlabeled data. In *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–I, April.

Weintraub, S., Rubin, N. P., and Mesulam, M.-M. (1990). Primary progressive aphasia: longitudinal course, neuropsychological profile, and language features. *Archives of neurology*, 47(12):1329–1335.

Williams, C., Thwaites, A., Buttery, P., Geertzen, J., Randall, B., Shafto, M., Devereux, B., and Tyler, L. (2010). The Cambridge cookie-theft corpus: A corpus of directed and spontaneous speech of brain-damaged patients and healthy individuals. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Yancheva, M. and Rudzicz, F. (2016). Vector-space topic models for detecting Alzheimer's disease. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2337–2346, Berlin, Germany, August. Association for Computational Linguistics.

Yang, B., Sun, J.-T., Wang, T., and Chen, Z. (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, pages 917–926, New York, NY, USA. ACM.

Yang, Y. (2001). A study of thresholding strategies for text categorization. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 137–145, New York, NY, USA. ACM.

Yimam, S. M., Gurevych, I., Eckart de Castilho, R., and Biemann, C. (2013). WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 1–6, Sofia, Bulgaria, August. Association for Computational Linguistics.

Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., and Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150 – 174.