Rare-Class Dialogue Act Tagging for Alzheimer's Disease Diagnosis

Shamila Nasreen[1], **Julian Hough**[1], **Matthew Purver**[1,2]
[1] Cognitive Science Group
School of Electronic Engineering and Computer Science
Queen Mary University of London, UK
[2] Jožef Stefan Institute, Ljubljana
{shamila.nasreen, j.hough, m.purver}@qmul.ac.uk

Abstract

Alzheimer's Disease (AD) is associated with many characteristic changes, not only in an individual's language, but also in the interactive patterns observed in dialogue. The most indicative changes of this latter kind tend to be associated with relatively rare dialogue acts (DAs), such as those involved in clarification exchanges and responses to particular kinds of questions. However, most existing work in DA tagging focuses on improving average performance, effectively prioritizing more frequent classes; it thus gives poor performance on these rarer classes and is not suited for application to AD analysis. In this paper, we investigate tagging specifically for rare class DAs, using a hierarchical BiLSTM model with various ways of incorporating information from previous utterances and DA tags in context. We show that this can give good performance for rare DA classes on both the general Switchboard corpus (SwDA) and an AD-specific conversational dataset, the Carolinas Conversation Collection (CCC); and that the tagger outputs then contribute useful information for distinguishing patients with and without AD.

1 Introduction

Natural Language Processing (NLP) has been applied to clinical health data for many purposes, including summarizing clinical notes, extracting specific elements from an unstructured medical record, and question-answer systems to interact with patients (Zahid et al., 2018; Velupillai et al., 2018; Demner-Fushman et al., 2009). Within this, one recent focus is on the use of NLP to diagnose the presence or extent of neurodegenerative cognitive impairment and/or monitor changes, based on patients' speech and language (see e.g. Roark et al., 2011), with much of this work focussing on dementia, primarily Alzheimer's Disease (AD) (see e.g. Orimaye et al., 2017). Most such approaches are based on features of the speaker's (or writer's) individual language, e.g. the complexity of vocabulary or syntax (see e.g. Fraser et al., 2016, for a comparison of a range of such features).

However, conditions such as AD also affect communication in interaction: AD patients display more conversational problems, often use terms that signal misunderstanding, and produce more requests for repair; while their conversational partners produce more elaboration or clarification (see e.g. Elsey et al., 2015). Closed (yes/no) questions are also asked more frequently of AD patients than open-ended wh-questions (Hamilton, 2005), and patients' ability to respond can vary with question type (Varela Suárez, 2018). Differences in dialogue act (DA) profiles might therefore add useful information for automatic diagnosis and monitoring of AD, and might also generalise better across languages than more lexically- or syntactically-based approaches: clarification and non-understanding signals seem to be quite general across languages and cultures (Dingemanse et al., 2015). However, while some computational studies have used interactional differences in AD diagnosis (see e.g. Luz et al., 2018; Mirheidari et al., 2019), these use models which are not interpretable in these DA terms, making it hard to provide useful output to clinical researchers, clinicians or carers.

Here, we therefore apply an explicit DA tagging approach to the problem, specifically looking for DAs that are characteristic of dementia, e.g. signals of non-understanding, requests for clarification, and particular types of questions and answers. Many of these are rare in natural dialogue, though; the *signal non-understanding* DA, for example, makes up only 0.1% of utterances in the Switchboard Corpus (Jurafsky et al., 1997). Standard DA tagging approaches, trained on average loss across all DA classes, therefore fail to give good performance.

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 290–300
July 29–31, 2021. ©2021 Association for Computational Linguistics

The main contributions of this paper are as follows:

- The adaptation of a hierarchical Bi-LSTM model to rare DA class tagging, modifying loss function, and the inclusion of contextual dependencies among DAs and utterances.
- Evaluation of the proposed method on two benchmark datasets, SwDA and CCC, achieving good performance: accuracy 88% with macro average F1 score 0.58 on SwDA, and accuracy 66% with F1 score 0.45 on CCC.
- Demonstration that these DAs can help distinguish between AD patients and Non-AD patients, achieving classification accuracy of 70% when used alone as unigram and bigram DA sequences, and 80% when combined with other interactional features.

2 Background

Interaction and AD diagnosis As explained above, AD patients display a number of characteristic interaction differences which can be characterised in terms of dialogue acts (DAs), including the rate of misunderstanding or non-understanding signals, requests for repair, elaboration, and clarification (Orange et al., 1996; Elsey et al., 2015), as well as yes/no-questions, wh-questions and choice questions and responses thereto (Hamilton, 2005; Gottlieb-Tanaka et al., 2003; Small and Perry, 2005; Varela Suárez, 2018). However, these studies, often based on Conversation Analysis (CA), give rich detail but are small-scale and/or qualitative. Some more quantitative corpus-based work makes similar observations: Nasreen et al. (2019) examine DA distributions in the Carolinas Conversation Collection (CCC, Pope and Davis, 2011), finding more signal-non-understanding, simple yes-answers and clarification requests in cognitively impaired patients' conversations.

Computational work that leverages these features is rare, however. Many diagnosis classification models include some signals associated with non-understanding (e.g. Fraser et al., 2016; Broderick et al., 2018) but only as part of large general language feature sets. One reason for this is that many studies use data that contains little interaction: the commonly used DementiaBank Pitt corpus, for example, contains conversations of a very one-sided nature. In a recent study, Farzana et al. (2020) developed an annotation scheme with 26 DAs based on ISO standard (Bunt, 2011) on DementiaBank data set to facilitate automated cognitive health screening from conversational interviews. They investigated phenomena like clarification request but some of the tags are specific to Cookie Theft Picture description task (Goodglass et al., 2001) and are not very general. Some recent work uses a more truly interactive approach: Luz et al. (2018) use a probabilistic graphical model to classify AD patients in the CCC corpus, although they use pauses and vocalisation times rather than any DA information; Mirheidari et al. (2019) include interactional features in a SVM classifier on Elsey et al. (2015)'s dataset, showing good accuracy, but use very specific features (e.g. "responding to neurologists' questions about memory problems") rather than more general DA tags. In contrast, our goal here is to investigate the use of general, well-known (but rare) DA classes.

Dialogue act (DA) tagging DA tagging has been approached using a range of machine learning techniques, starting with early work using Hidden Markov Models to capture the intuition that key information lies in both the sequences of words within utterances and the sequence of DAs across utterances (Stolcke et al., 2000). Improvements have been gained by using Conditional Random Fields (Zimmermann, 2009), cue phrase models (Webb et al., 2005), joint classification and segmentation (Ang et al., 2005), and more recently neural networks including Recurrent Neural Networks (RNNs) (Kalchbrenner and Blunsom, 2013; Ortega and Vu, 2017) and Convolutional Neural Networks (CNNs) (Lee and Dernoncourt, 2016). Most recent work sticks with Stolcke et al. (2000)'s original intuition to include contextual information (preceding utterances and their DA roles help predict the current utterance), often via hierarchical models where the higher layers capture DA/utterance sequence information; see e.g. (Raheja and Tetreault, 2019)'s use of a CRF above dialogue-level and utterance-level BiLSTMs, achieving state-of-the-art accuracy of 82.9% on the standard SwDA corpus. However, variants exist: Bothe et al. (2018), for example, consider only a limited number of preceding utterances as a context within a RNN, rather than the full sequence, accuracy is reduced to 77.34% on SwDA but their model, in using only limited preceding context (rather than assuming knowledge of future utterances) is suitable for incremental online settings.

Rare DA classes All these approaches, however, train and evaluate their models assuming that the goal is average performance over a general DA tagset, usually the 42-tag SwDA DAMSL scheme (Stolcke et al., 2000). Some use fewer classes — Fuscone et al. (2020) use 3 dominating DA classes *statement*, *opinion*, and *backchannel*; Ramacandran (2013) use an 18-tag DAMSL subset; Sridhar et al. (2009) group the 42 classes into 7 common classes and one 'other' category based on frequency — but all of these focus on the most common tags. In contrast, we are interested in the rare classes useful for dementia analysis, following the clinical CA work described above; we give a full list of these classes of interest in Section 4.1 (see Table 1). Few studies give details of accuracy on these rarer classes; but Raheja and Tetreault (2019), despite achieving 82.9% accuracy overall, show accuracy of only c.25% for *br* (*signal-non-understanding*, which makes up only 0.1% of SwDA utterances), c.30% for *b^m* (*repeat-phrase*, 0.3% of utterances), c.20% for *qy* (*yes-no-question*, 2%), and <5% for both *qw* (*wh-question*, 1%) and *b* (*backchannel*, a relatively common but important tag).

3 Proposed Approach

Here, then, our purpose is to improve DA tagging accuracy for the specific DA classes of interest in AD diagnosis, including specific types of questions, answers and misunderstanding signals, most of which are relatively rare. For this purpose, we use a context-based hierarchical BiLSTM model with attention, to capture relations at the word, utterance and DA level and leverage utterance DA/context information. To maintain the ability to use our model in an online setting, we use only utterances from the preceding (left) context, not the following (right) context. We perform DA tagging experiments on two corpora, one general and one AD-specific, to compare a range of models:

- A baseline model using the word embeddings as text features, without any context information;
- A hierarchical BiLSTM model using word embeddings and previous utterance representations from context;
- A hierarchical BiLSTM model using word embeddings, previous utterance representations and previous predicted DA tags from context.

3.1 Model Representation

Formally, we model each dialogue conversation D as a sequence of utterances $U = \{U_1, U_2, U_3, ..., U_n\}$ paired with a sequence of DA labels $Y = \{da_1, da_2, da_3, ..., da_n\}$; each utterance $U_t \in U$ is a sequence of words $U_t = \{w_t^1, w_t^2, ..., w_t^m\}$.

Figure 1 shows the overall architecture of our model in which U_t represents the current utterance and U_{t-1} represents the previous utterance. We use word embeddings to extract the lexical feature representations from the transcripts, converting the utterances from word sequences into sequences of word vectors. We compared the use of randomly initialised embeddings, GloVe pretrained embeddings (Pennington et al., 2014), GloVe embeddings trained on SwDA and CCC corpus, and ELMo embeddings (Peters et al., 2018).

This word representation layer feeds into a BiLSTM, producing a representation of an utterance as a sequence of hidden vectors $h_t = \{h_t^1, h_t^2, ..., h_t^m\}$. We use an attention mechanism to weight these and aggregate them into a single utterance representation, an attention vector c_t is representing the whole utterance U_t. We then concatenate the vector for the current utterance c_t with various combinations of information from previous context: the previous utterance vector c_{t-1}, previous DA (da_{t-1}) (gold-standard or predicted, see Section 4), and their preceding neighbours c_{t-2}, da_{t-2}. These concatenated vectors are then encoded by a second LSTM (here, we use a unidirectional left-to-right LSTM, rather than bidirectional, to stay compatible with utterance-by-utterance online processing); the resulting sequence of hidden vectors H=$\{H_1, H_2, ..., H_n\}$ is then used to predict da_t, the DA label of the current utterance U_t.

4 Experiments

4.1 DA filtering

To keep our approach as domain- and dataset-general as possible, we start with the standard DAMSL tagset (Stolcke et al., 2000) and adapt it. Based on the clinical studies described in Section 2, we keep 17 specific DA tags of interest from DAMSL; split 2 of them each into 2 sub-categories; and collapse all other tags into a single **other** tag, giving a total of 20 tags. The two new DA tags are **clarification-request** *(qc)* and **statement-answer** *(sa)*: clarification-request *(qc)* is a sub-category of *signal-non-understanding (br)* which

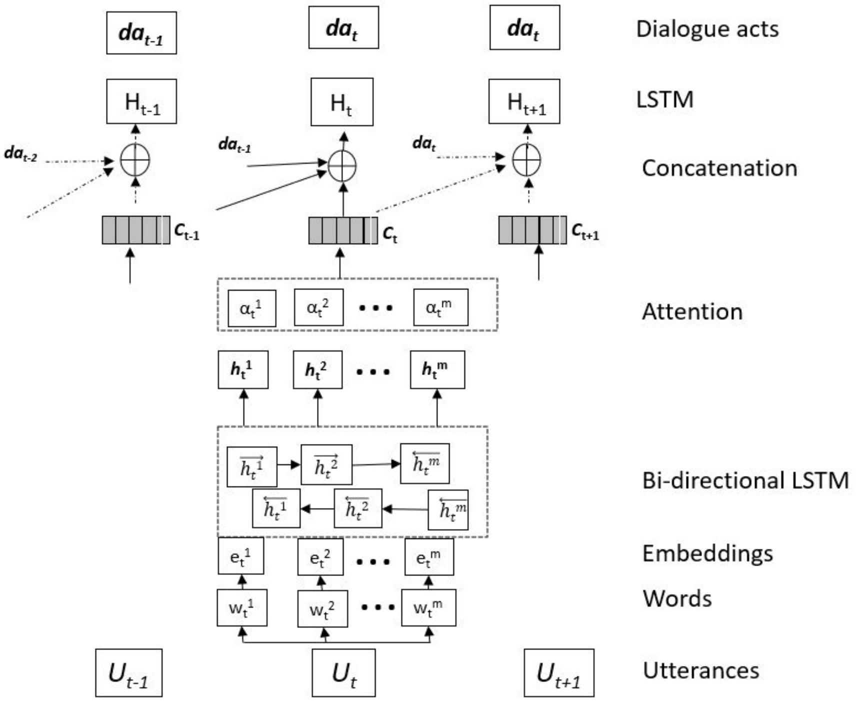

Figure 1: Model architecture for DA classification with one utterance and one DA as context.

requests more specific information (see e.g. Purver et al., 2001; Rodríguez and Schlangen, 2004); while *statement-answer (sa)* is a sub-category of *declarative-statement (sd)* used as an answer to a wh-question *(qw)*, open-question *(qo)* or or-question *(qr)*. The full tagset[1] is shown in Table 1.

4.2 Datasets

We evaluated our model on two datasets. First, the standard **Switchboard Corpus (SwDA)** transcripts, a corpus of 1155 five-minute two-speaker telephone dialogues, containing 205K utterances in total (Jurafsky et al., 1997). Second, the **Carolinas Conversation Collection (CCC)**[2] transcripts, a corpus of transcribed audio about the health of people over 65 years of age in natural conversations (Pope and Davis, 2011). The CCC is a systematic collection of two cohorts: one contains conversations of 125 patients with AD who spoke twice at least with a researcher; the other contains conversations from elderly persons with different medical conditions, recorded twice a year, once with a researcher and once with a community person in the home or community settings. Each patient is interviewed by a different interviewer. The CCC includes some uniform questions that are collection-specific for people specific to health conditions, diseases, and cognitively-impaired speakers with dementia. It is transcribed but not annotated with DA tags. Access to the data was granted after ethical review by the both Queen Mary University of London (via QMERC2019/04 dated:25-04-2019) and MUSC.

4.2.1 Manually Tagged Annotations

We performed manual annotation of the CCC corpus with DA tags using the SwDA-derived tagset of Section 4.1 above. We annotated 20 conversations with 10 Non-AD patients from one cohort, and 10 conversations with AD patients from the other, giving a total of 30 conversations [3]. Comparing three annotators on one sample conversation, we achieved an inter-rater agreement of 0.844.

For the SwDA corpus, we reduced the original 42-tag labels to our reduced tagset. This required manual re-tagging of some *signal non-understanding* utterances with the new subcate-

[1] The annotation guidelines are available from https://osf.io/8w9z2/?view_only=ee08242870f24ae7ab6754ddf9a0176a.

[2] https://carolinaconversations.musc.edu/

[3] The annotations are available for research community for further followup work and can be useful after getting access to CCC dataset: https://osf.io/8w9z2/?view_only=ee08242870f24ae7ab6754ddf9a0176a

Tagset	Label	Example	Percentage in SWDA
Yes-No Question	qy	Did you go anywhere today?	2%
Wh-Question	qw	When do you have any time to do your homework?	1%
Declarative Yes-No Question	qy^d	You have two kids?	1%
Declarative Wh-Question	qw^d	Doing what?	<0.1%
Or Question	qr	Did he um, keep him or did he throw him back?	0.1%
Tag Question	^g	But they're pretty aren't they?	<0.1%
Open ended question	qo	And uh -how do you think -that work helps you?	0.3%
Clarification Question	qc	Next Tuesday?	-
Signal Non-understanding	br	Pardon?	0.1%
Backchannel in question form	bh	Really?	1%
Yes answer	ny	Yeah.	1%
Yes- plus expansion	ny^e	Yeah, but they're .	0.4%
Affirmative non-yes answer	na	Oh I think so. [laughs]?	0.4%
No answer	nn	No	1%
Negative non-no answers	nn^e	No, I belonged to the Methodist church.	0.1%
Other answers	no	I, I don't know.	1%
Statement answer	sa	Popcorn shrimp and it was leftover from yesterday.	-
Backchannel(continuer)	b	Uh-huh	19%
Repeat phrase	b^m	Ahh, Corn Bread.	0.3%
Other	Other	*(everything else)*	71.1%

Table 1: Rare class DA tagset with their Labels and Example.

Class	Prec.	Rec.	F1
sa	1	0.83	0.90
sd	0.86	1	0.92

Table 2: Prediction score for Rule-based classification,

Dataset	SwDA	CCC
Transcripts	1115	30
Total utterances	142022	5082
Training utterances	111356	-
Test utterances	27840	5082

Table 3: Both datasets with number of utterances.

gory *clarification-request*, and similarly re-tagging some *declarative statement* utterances as *statement answer* (*sa*). The latter could be achieved semi-automatically, as the new *statement answer* category can only apply in response to *qw*, *qr*, and *qo* questions: we took 8 conversations from the SwDA corpus containing 27 questions (*qw, qr, qo*), and manually re-tagged their answers from *sd* to *sa*. From this, we then built a rule-based classifier to derive simple rules for conversion of *sd* statements to *sa* tags, applied to the rest of the corpus. The accuracy of this rule-based classifier is reported in Table 2. We then used the standard train/test split for SwDA; we train only on SwDA, keeping CCC purely as a test set. Table 3 shows the statistics from both corpora.

4.3 Implementation and metrics

We performed a grid search for hyperparameter tuning, changing one hyperparameter at a time. We trained our model using ADAM (Kingma and Ba, 2014) with a learning rate of 0.01 and used categorical cross-entropy as the loss function for the multi-class outcomes. As the classes in our data are highly imbalanced, we use a class-weighted objective function to prevent over-prioritising more common classes; use scikit-learn's StratifiedShuffleSplit (a merge of StratifiedKFold and ShuffleSplit) to preserve the percentage of each class in each fold. Embedding size was set to 100 dimensions for both simple word embeddings and GloVe pretrained embeddings, with 1024 dimensions for ELMo embeddings. We report accuracy, macro-average precision (Prec.), recall (Rec.), and F1 score for multi-class classification. We choose macro-average measures as our data is highly imbalanced and we are particularly interested in the rare DA classes.

Baseline Model We define our base model for single utterance classifications at the sentence level without including any contextual utterance or DA information.

5 Results

Table 4 shows the performance of our baseline model (without context) and the proposed models with a range of context settings: with one, two, and three previous utterances and previous DA tags as context. Our best baseline model (using ELMo embeddings) yields a macro-averaged F1 score of 0.46 on the SwDA test set and 0.34 on the CCC test set. Results are improved by adding contextual in-

Context	Embedding	SwDA test set				CCC test set			
		Prec.	**Rec.**	**F1**	**Acc.**	**Prec.**	**Rec.**	**F1**	**Acc.**
No Context (Baseline)	No Pretrain	0.42	0.47	0.42	0.79	0.33	0.34	0.31	0.50
	Glove	0.44	0.46	0.44	0.83	0.38	0.36	0.32	0.53
	ELMo	0.45	0.55	0.46	0.80	0.37	0.37	0.34	0.52
1 utt only	No Pretrain	0.45	0.57	0.49	0.81	0.44	0.44	0.41	0.55
	Glove	0.48	0.57	0.51	0.83	0.46	0.48	0.43	0.57
	ELMo	0.43	0.54	0.45	0.78	0.40	0.38	0.35	0.52
1 utt & 1 DA	No Pretrain	0.52	0.62	0.56	0.87	0.49	0.45	0.44	0.62
	Glove	**0.55**	**0.62**	**0.57**	**0.88**	**0.48**	**0.47**	**0.45**	**0.62**
	Glove Swda-CCC	**0.57**	**0.61**	**0.58**	**0.88**	**0.51**	**0.48**	**0.45**	**0.66**
	Glove (SP info.)	0.54	0.64	0.57	0.87	0.46	0.46	0.43	0.64
	ELMo	**0.55**	**0.64**	**0.58**	**0.88**	0.47	0.43	0.40	0.62
2 utt only	No Pretrain	0.46	0.53	0.49	0.82	0.37	0.36	0.33	0.53
	Glove	0.48	0.57	0.50	0.82	0.44	0.43	0.40	0.55
	ELMo	0.42	0.45	0.40	0.81	0.40	0.38	0.33	0.51
2 utt & 2 DAs	No Pretrain	0.52	0.62	0.56	0.87	0.44	0.45	0.42	0.63
	Glove	0.56	0.59	0.57	0.88	0.48	0.46	0.43	0.69
	ELMo	0.59	0.59	0.56	0.88	0.49	0.43	0.42	0.63
3 utt only	No Pretrain	0.35	0.49	0.40	0.77	0.42	0.33	0.33	0.49
	Glove	0.32	0.43	0.35	0.79	0.35	0.31	0.3	0.51
	ELMo	0.44	0.45	0.39	0.76	0.33	0.38	0.3	0.52
3 utt & 3 DAs	No Pretrain	0.51	0.59	0.54	0.87	0.39	0.41	0.37	0.60
	Glove	0.52	0.64	0.56	0.87	0.44	0.45	0.41	0.61
	ELMo	0.51	0.53	0.48	0.88	0.41	0.43	0.36	0.60

Table 4: Accuracy, macro-average precision, recall, and F1 score for different contexts with different word embeddings on **SwDA test set** and **CCC test set**.

formation from previous utterances and further improved by adding previous DA labels. Our model achieved a macro-average F1 score of 0.51 with only one utterance as context, further improved by to 0.57 by considering the previous utterance DA label (SwDA corpus, GloVe embeddings). With ELMo embeddings, F1 score is lower than GloVe for one utterance context (0.45 F1) but increases more when adding the DA information, giving our best performance **(Rec.:0.64, F1: 0.58, Acc.: 0.88)** on SwDA. Transferring the model learned on SwDA to the AD-specific CCC corpus also gives its best result in this setting: we obtain our best macro F1 score of 0.45 on CCC when using one preceding utterance and one DA as context with GloVe embeddings. Using GloVe embeddings trained on the SwDA and CCC data perhaps gives slight improvements over the standard pre-trained GloVe, but they are small (Table 4).

We also experimented with different variants of including speaker identity information (e.g. by concatenating speaker ID with DA history); this did not improve results, so we report it only for the best context setting as illustration. Overall, these results suggest that the single immediately preceding utterance and DA label have the largest impact on performance: including more context history does not help, and using preceding DAs as well as preceding utterances as context is more effective than using utterances alone. Overall, all the methods using context yield significant improvement over the baseline.

Model	DA	Prec.	Rec.	F1
1 utt & 1 DA	G	0.55	0.62	0.57
1 utt & 1 DA	P	0.51	0.54	0.49
2 utt & 2 DAs	G	0.56	0.59	0.57
2 utt & 2 DAs	P	0.51	0.52	0.48
3 utt & 3 DAs	G	0.52	0.64	0.56
3 utt & 3 DAs	P	0.58	0.49	0.51

Table 5: Comparison of models using gold-standard (G) DAs label as context vs using predicted (P) DAs as context on SwDA test set. These reported results are macro-averages.

Table 4 uses gold-standard contextual DA tag information; this raises the question of whether adding DA information would be less effective when using predictions. We therefore compared the use of predicted (P) DA labels vs. gold-standard (G) DA labels as context when testing, shown in Table 5. We achieve better performance when using the gold-standard labels in both training and testing, as expected; on the other hand, when training on gold-standard labels but using previously predicted DAs as context during testing — a more realistic approach in real-time systems — we achieve reasonable performance which improves as the context window increases, suggesting that further improvements may be gained by using more predicted DA labels as context.

Our interest, of course, is not in macro-average figures but in predicting the distribution over the individual DA classes. We therefore, examine the class-wise prediction scores, showing a selection of classes in Figure 2. We note that performance exceeds that of Raheja and Tetreault (2019) (see Section 2) by a very large margin in all cases. Class-wise results for each class in our tagset can be found in supplementary materials.

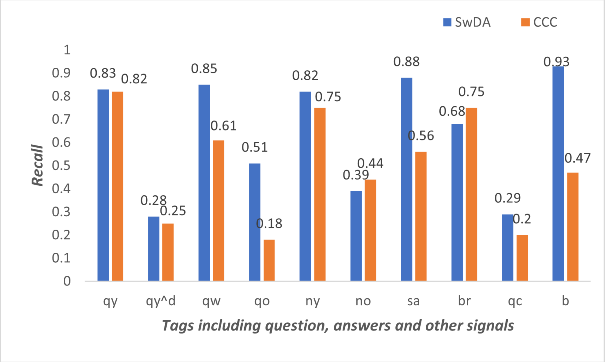

Figure 2: Comparison of class-wise recall for individual DA tags on both SwDA and CCC datasets.

Error Analysis We conducted an error analysis to closely look into the lower performance of the model for some DA classes. We observed poor recall scores for *qwˆd* in both corpora and for *qo* questions in CCC. Most of the *qo* and *qwˆd* questions are mislabeled with *qw* tag or *other* tag. This is somewhat reasonable, as linguistically the utterances of these classes are quite similar, although *qw* and *qwˆd* express more specific questions whereas *qo* utterances tend to be general, and they share many syntactic cues which can easily confuse the model. A few *qwˆd* questions were also misclassified as either *qyˆd* or *qy*.

Clarification request *(qc)* recall values are low in both datasets; upon analysis, we found that *qc* is often confused with signal non-understanding *(br)* and wh-questions *(qw)*. For example, *qc* utterances with forms such as 'Youre now in what?', 'You must be what?', 'being what?', 'what's that?', although requesting clarification in context, are understandably easy to mislabel as *qw*. Encouragingly, including utterance/DA context improved these results. Recall scores for backchannels *(b)* are high for SwDA but lower for CCC. One possible reason could be the different transcription protocols in the two datasets: some transcribers use 'yeah', 'yup' while others can use the standard form 'yes' to represent a backchannel. Some surface forms of backchannels are also present in the CCC dataset but did not occur in SwDA, and are thus misclassified when testing on CCC.

We further analyzed the effect of adding utterance/DA context on individual DA classes, with results shown in Figure 3 and Figure 4. Yes-answer (*ny*) recall improved from 0.22 to 0.58 when including only one preceding utterance, and is further improved to 0.75 by adding the previous DA label. A simple statement 'yes' can be an answer or a backchannel (amongst others); the information that the previous DA label may be a yes-no question (*qy*) will help in distinguishing the two.

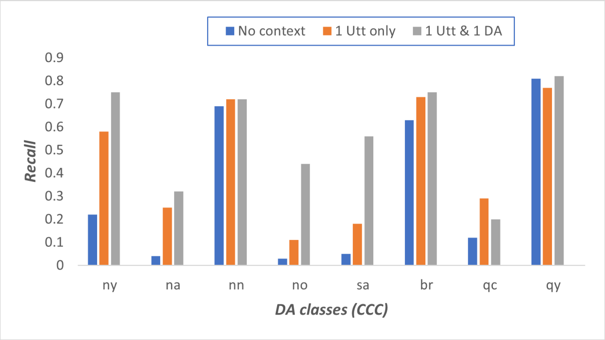

Figure 3: Effect of including context on DA prediction on CCC test set.

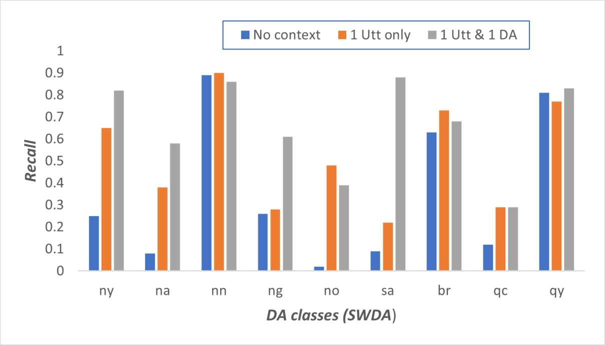

Figure 4: Effect of including context on DA prediction on SwDA test set.

6 Using DA tag outputs for AD diagnosis

Finally, we performed an initial investigation of the use of our tagger outputs in the eventually intended downstream task: the usage of these DA patterns to diagnose AD. We treat this here as a classification task, distinguishing between dialogues involving AD patients and Non-AD patients (similar age controls) in the CCC corpus. As an initial experiment, we use the DA classes (shown in Table 6) investigated in our experiments above as features within a linear SVM classifier, and report results in Table 7. We tested the use of the DA classes both as unigrams (*f1*) and as bigrams (*f2*) to capture characteristic local DA sequences. For this experiment, we only used bigram sequences containing the meaning-coordination *qc* and *br* DAs in patient *(P)* utterances, preceded by question DAs from the interviewer *(I)*. We also computed two aggregate features from these DAs as proxies for levels of patient confusion (*f3*): question_ratio (how many questions asked by the patient (*P*) out of total utterances spoken by *P*) and confusion_ratio (ratio of total *br* & *qc* to the total questions asked by *P*). Question_ratios were previously used by Khodabakhsh et al. (2015) in AD identification, considering question words such as 'what', 'which' etc. as a mark of confusion or request for further details. Here, we replicate that as question_ratio, and add the more specific use of *qc* and *br* tags as confusion_ratio. We further experiment with other useful interactional features (*f4*) such as normalized turn lengths, an average number of words per minute (as used by Luz et al. (2018) for AD prediction), turn switches per minute, and number of overlaps. Overlaps represent the number of segments spoken simultaneously by both speakers, with the intuition that these may be attributed to speech initiation difficulties.

Features	**Type (Total)**	**Details**
f1	Unigrams (36)	unigram DAs such as: *P_qy, P_ny, P_br, P_na, P_sa, I_qo, I_qw, I_b, I_qy*
f2	Bigrams (17)	bigram DAs sequences such as: *I_qw–P_br*, *I_qo–P_ny*, *I_sa–P_qc* *I_qw–P_qc*, *I_qwˆd–P_qc*
f3	Confusion (2)	question_ratio, confusion_ratio
f4	Others (4)	other features from dialogues includes: normalized turn duration, Avg number of words per minute, turn switches per minute, number of overlaps

Table 6: Different features for AD classification task.

We achieved an accuracy of 0.65 with only unigrams, 0.70 when including bigram sequences and confusion features, over a random baseline[4] of 0.50. Combining these with other interactional features improved the results to an overall accuracy of 0.80. We conclude that our rare-class tagger provides suitable accuracy to be used in future work in AD diagnosis and monitoring.

Model	**Features**	**class**	**Prec.**	**Rec.**	**F1**	**Acc.**
Random	-	AD	0.50	0.50	0.50	0.50
(baseline)		Non-AD	0.50	0.50	0.50	
SVM	*f1*	AD	0.67	0.60	0.63	0.65
		Non-AD	0.67	0.70	0.67	
SVM	*f1,f2,f3*	AD	0.68	0.80	0.73	0.70
		Non-AD	0.75	0.60	0.67	
SVM	*f1,f2,f3,f4*	AD	0.75	0.90	0.82	0.80
		Non-AD	0.87	0.70	0.78	

Table 7: Results on the AD classification task on CCC data.

7 Conclusion

This work has presented a DA tagger (a hierarchical BiLSTM model) with a context-based learning approach for the classification of rare DAs including clarification requests, non-understanding signals, questions, and responses. By using suitable choices of embeddings and the inclusion of contextual history, together with a weighted cost function, we achieve good performance on these rare classes. For SwDA, our model achieved F1 of 0.58 and recall of 0.64 when using the immediate preceding utterance and DA label, compared to F1 of 0.46, recall of 0.55 without context. We found that while gold-standard DA information from context gives better performance, the performance using predicted labels can be improved by using longer contextual sequences.

The resulting DA tagger utilizes only minimal context of a few preceding utterances and DAs, rather than the whole conversation, and thus is suitable for dialogue systems in real-time, due to the left-to-right, incremental nature of dialogue. Existing models which take into account the whole conversation can achieve overall higher accuracy on the general DA tagging task, and so might be expected to improve our rare-class task as well, but require information about future utterances (Li et al., 2018; Raheja and Tetreault, 2019).

Its rare-class DA outputs show good potential as features to distinguish between AD and Non-AD patients in interaction, suggesting that they can be useful within tools to aid in diagnosis while provid-

[4] An alternative, stronger baseline could be the use of a standard DA tagger trained on the general 42-class tagset, to isolate the improvement gained specifically by our focus on rare class DAs. Unfortunately this is not currently possible, as the CCC corpus has no transcripts tagged in this way.

ing useful, interpretable information about interaction structure, mutual understanding, and question-answering behavior. Phenomena such as clarification requests and signals of non-understanding seem to be quite general across languages and cultures (Dingemanse et al., 2015) and we would expect these sorts of conversational features to be more language- and domain-independent than approaches based on vocabulary, syntax, etc for AD diagnosis. We note, however, that one limitation of this study is that the AD patients in the CCC dataset are all older patients with already diagnosed dementia, and can thus only allow us to observe patterns associated with AD at a relatively advanced stage, and not directly tell us whether these extend to early-stage diagnosis.

In future, we will improve the performance of our rare class DA tagger with the inclusion of acoustic features from speech data. We also hope to explore more informative DA sequences, including other bigram and trigram sequences, while retaining the interpretable nature of the model overall.

Acknowledgments

Purver is partially supported by the EPSRC under grant EP/S033564/1, and by the European Union's Horizon 2020 programme under grant agreements 769661 (SAAM, Supporting Active Ageing through Multimodal coaching) and 825153 (EMBEDDIA, Cross-Lingual Embeddings for Less-Represented Languages in European News Media).

References

Jeremy Ang, Yang Liu, and Elizabeth Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 1:I/1061–I/1064 Vol. 1.

Chandrakant Bothe, Cornelius Weber, Sven Magg, and Stefan Wermter. 2018. A context-based approach for dialogue act recognition using simple recurrent neural networks. *arXiv preprint arXiv:1805.06280*.

Brianna Marlene Broderick, Si Long Tou, and Emily Mower Provost. 2018. TD-P-014: Cogid: A speech recognition tool for early detection of Alzheimer's disease. *Alzheimer's and Dementia*, 14(7S).

Harry Bunt. 2011. Multifunctionality in dialogue. *Computer Speech & Language*, 25(2):222–245.

Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. 2009. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772.

Mark Dingemanse, Seán G Roberts, Julija Baranova, Joe Blythe, Paul Drew, Simeon Floyd, Rosa S Gisladottir, Kobin H Kendrick, Stephen C Levinson, Elizabeth Manrique, et al. 2015. Universal principles in the repair of communication problems. *PloS one*, 10(9):e0136100.

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, 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.

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

Simone Fuscone, Benoit Favre, and Laurent Prévot. 2020. The contribution of dialogue act labels for convergence studies in naturalconversations.

Harold Goodglass, Edith Kaplan, Sandra Weintraub, and Barbara Barresi. 2001. The boston diagnostic aphasia examination.

Dalia Gottlieb-Tanaka, Jeff Small, and Annalee Yassi. 2003. A programme of creative expression activities for seniors with dementia. *Dementia*, 2(1):127–133.

Heidi Ehernberger Hamilton. 2005. *Conversations with an Alzheimer's patient: An interactional sociolinguistic study*. Cambridge University Press.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *URL http://web. stanford. edu˜ jurafsky/ws97/manual. august1. html*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *arXiv preprint arXiv:1306.3584*.

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):1–15.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Ruizhe Li, Chenghua Lin, Matthew Collinson, Xiao Li, and Guanyi Chen. 2018. A dual-attention hierarchical recurrent neural network for dialogue act classification. *arXiv preprint arXiv:1810.09154*.

Saturnino Luz, Sofia de la Fuente, and Pierre Albert. 2018. A method for analysis of patient speech in dialogue for dementia detection. In *Proceedings of the LREC 2018 Workshop Resources and Processing of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments (RaPID-2)*.

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, 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. SEMDIAL, London, United Kingdom (Sep 2019), http://semdial.org/anthology/Z19-Nasreen semdial*, volume 13.

John 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.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. Predicting probable Alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18(1):34.

Daniel Ortega and Ngoc Thang Vu. 2017. Neural-based context representation learning for dialog act classification. *arXiv preprint arXiv:1708.02561*.

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

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

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

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the means for clarification in dialogue. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Nithin Ramacandran. 2013. Dialogue act detection from human-human spoken conversations. *International Journal of Computer Applications*, 67(5).

Brian Roark, Margaret Mitchell, John-Paul Hosom, Kristy Hollingshead, and Jeffrey Kaye. 2011. Spoken language derived measures for detecting mild cognitive impairment. *IEEE transactions on audio, speech, and language processing*, 19(7):2081–2090.

Kepa Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in German task-oriented spoken dialogues. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue (Catalog)*, pages 101–108, Barcelona.

Jeff A Small and JoAnn Perry. 2005. Do you remember? how caregivers question their spouses who have Alzheimer's disease and the impact on communication. *Journal of Speech, Language, and Hearing Research*, 48(1):125–136.

Vivek Kumar Rangarajan Sridhar, Srinivas Bangalore, and Shrikanth Narayanan. 2009. Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, 23(4):407–422.

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):339–373.

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

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, et al. 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.

Nick Webb, Mark Hepple, and Yorick Wilks. 2005. Dialogue act classification based on intra-utterance features. In *Proceedings of the AAAI Workshop on Spoken Language Understanding*, volume 4, page 5. Citeseer.

MAH Zahid, Ankush Mittal, Ramesh Chandra Joshi, and G Atluri. 2018. Cliniqa: A machine intelligence based clinical question answering system. *arXiv preprint arXiv:1805.05927*.

Matthias Zimmermann. 2009. Joint segmentation and classification of dialog acts using conditional random fields. In *Tenth Annual Conference of the International Speech Communication Association*.