INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Diagnosing people with dementia using automatic conversation analysis

Bahman Mirheidari[1], *Daniel Blackburn*[2], *Markus Reuber*[3], *Traci Walker*[4], *and Heidi Christensen*[1]

[1]Department of Computer Science, University of Sheffield, Sheffield, UK
[2]Sheffield Institute for Translational Neuroscience (SITraN), University of Sheffield, Sheffield, UK
[3]Academic Neurology Unit, University of Sheffield, Royal Hallamshire Hospital, Sheffield, UK
[4]Department of Human Communication Sciences, University of Sheffield, Sheffield, UK
{bmirheidari2,d.blackburn,m.reuber,traci.walker,heidi.christensen}@sheffield.ac.uk

Abstract

A recent study using Conversation Analysis (CA) has demonstrated that communication problems may be picked up during conversations between patients and neurologists, and that this can be used to differentiate between patients with (progressive neurodegenerative dementia) ND and those with (nonprogressive) functional memory disorders (FMD). This paper presents a novel automatic method for transcribing such conversations and extracting CA-style features. A range of acoustic, syntactic, semantic and visual features were automatically extracted and used to train a set of classifiers. In a proof-of-principle style study, using data recording during real neurologist-patient consultations, we demonstrate that automatically extracting CA-style features gives a classification accuracy of 95 % when using verbatim transcripts. Replacing those transcripts with automatic speech recognition transcripts, we obtain a classification accuracy of 79 % which improves to 90 % when feature selection is applied. This is a first and encouraging step towards replacing inaccurate, potentially stressful cognitive tests with a test based on monitoring conversation capabilities that could be conducted in e.g. the privacy of the patient's own home.

Index Terms: speech recognition, machine learning, conversation analysis, dementia

1. Introduction

The sensitivity and specificity of current screening procedures for possible dementia in primary care is suboptimal: patients with a high level of cognitive functioning may not be identified as being at risk of dementia whereas those with psychiatric causes of memory problems, but no evidence of dementia, are referred to specialist memory clinics although they could have been treated in primary care. The latter patient group currently makes up 50% of referrals to neurology-led secondary care memory services in the UK [1, 2]. An automatic, diagnostic tool for the early detection of dementia is therefore highly desirable. It would reduce pressure on services and enable doctors to provide appropriate care and reassurance to patients whose memory problems are not related to dementia.

Dementia is a disorder of the brain which can be caused by a number of diseases such as Alzheimer's Disease. One of the most widely know symptoms is problems with memory, and from early on this will also affect a person's language and their ability to conduct a normal conversation – something neurologists will often notice as they start the routine history-taking part of their assessment consultation.

A recent study applied conversation analysis (CA) to such doctor-patient interactions and found a set of 6 language characteristics that could be used to distinguish between patients with neurodegenerative dementia (ND) and patients with Functional Memory Disorder (FMD) (not dementia-related) [3, 4]. The study showed promising results in terms of diagnostic power, but relied on manual CA for discovering the interaction patterns in the conversation; this involves a number of steps including audio recording, transcribing the encounters and carrying out a qualitative analysis by a trained expert. It is thus prohibitively expensive and time consuming and not feasible for large-scale use. This paper presents work towards an *automatic* CA-based dementia detection system where dedicated speech technology software is used to analyse the audio-recorded interactions.

Automatic CA is an emerging and challenging area of research with some promising results (e.g. [5, 6]). It typically involves a number of technologies to automate the above steps including automatic speech recognition (ASR), speaker diarization (*"who's speaking when"*) and some automatic speech understanding.

Using CA in the framework of diagnosing dementia is novel, however, related work on using machine learning techniques to identify signs of dementia in patients speech and language exists. Researchers have tried to combine various types of utterance-level features to discriminate people with dementia from healthy controls – an easier task than the current ND/FMD discrimination task. Lopez and *et al.* investigated acoustic features such as duration, time domain, frequency domain, and the Higuchi Fractal Dimension from a database called AZTIAHO of multilingual recordings of spontaneous speech of 50 healthy adults and 20 Alzheimer's patients [7, 8]. Jerrold *et al.* [9] mixed up half of the ASR outputs with half of human transcription of spontaneous speech to extract acoustic and lexical features used for classifying 48 participants with different types of dementia and a healthy control group. The classification accuracy amongst all types of subjects was 61%, while the binary classification between AD and healthy controls rose to 88% accuracy.

Thomas *et al.* [10] extracted several syntactic and semantic features to achieve a 95 % accuracy for a binary classification task differentiating between patients with severe dementia and healthy controls. Recent research [11, 12], have used the DementiaBank corpus to predict the score of the MMSE assessment of the patients over time. They extracted a wide range of features (over 477 lexicon syntactic, acoustic, and semantic) and selected the 40 most informative features, reporting accuracy over 92% in distinguishing AD patients from the healthy control group. However they worked with manually produced transcripts of the audio files and it is unclear how their results would be affected by using ASR.

http://dx.doi.org/10.21437/Interspeech.2016-857

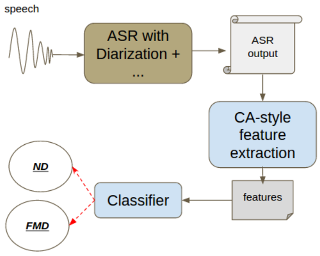

Figure 1: *Block diagram of the dementia detection system.*

In this paper, we describe our results of replicating, as closely as possible, the findings of Elsey *et al.* [3] in a proof-of-principle style study. We investigate the effects of automatising the CA-style features extraction as well as the effect of coupling that with erroneous ASR-produced transcripts. Section 2 describes the system, section 3 describes the experimental setup and sections 4 and 5 discuss the results and conclusions.

2. Dementia detection system

Figure 1 shows the block diagram for the proposed automatic dementia detection system. First, given the audio recording of the conversation, the ASR and speaker diarization module outputs transcripts with time stamped word IDs as well as speaker turns. This is passed to the feature extraction unit where CA-style features are extracted. Finally, the features are sent to a machine learning classifier trained to discriminate between ND and FMD conversations. As the main aim of the study has been to demonstrate the feasibility of creating an automatic method of the diagnostic profiling in Elsey *et al.*, only features directly related to those CA profile characteristics, identified in the original study, has been considered.

3. Experimental setup

3.1. Data

The data used in Elsey *et al.* [3] comprises of a total of 30 audio recordings and associated manual CA annotations of patient-neurologist conversations. For this study we have been able to use an extended data set comprising of 39 conversations (21 FMD and 18 ND). Patients were encouraged to bring an accompanying person, so many conversations have three participants. The neurologists were instructed to try to stick to a predefined set of questions constructed so as to reveal the typical signs of impairments in the conversation. Several categories of questions were included:

- Closed questions needing long-term memory recall of personal details the person is meant to know (e.g, ”How old where you when you left school?”).
- Compound questions (e.g., ”Why have you come here today, and what are your expectations?”). People with dementia will find it difficult to remember to answer both parts.
- Open-ended questions like ”What did you do after you left school?”.
- Questions related to the memory concern, like ”Who’s the most worried about your memory?” (for ND patients it tends to be other members of the family who are worried about the patient) and ”Tell me about the last time you had a problem with your memory”, which FMD patients find easier to answer.

As the data was not initially recorded with the aim of applying speech recognition, little effort has was made to reduce background noise and acoustic interference, and for many of the recordings the microphone placement has been relatively *ad hoc* (often being placed closer to the neurologist than the patient). In addition, the speech itself is very challenging with a high percent of overlapping speech segments - on occasion even the professional transcribers have not been able to transcribe the material. For this initial study it was therefore chosen to only include segments with non-overlapping speech.

3.2. Feature extraction

The primary objective in Elsey *et al.* were to define a set of characteristics[1] that would enable a *diagnostic profile* to be drawn up for a patient; a total of six such characteristics were defined and are described in Table 1 along with descriptions of the corresponding automatic features that are proposed for the current study. The 6 profiling characteristics are: (Role of) ”accompanying person” (F1), ”responding to neurologists’ questions about memory problems” (F2), ”Patient recall of recent memory failure” (F3),”responding to compound questions” (F4), ”inability to answer” (F5), ”and patient’s elaborations and length of turn” (F6).

A total of 22 features were defined to replicate as closely as possible the original, manually extracted features in Elsey *et al.*. Most of these features are extracted individually for each of the two/three conversation participants and the features are named accordingly using prefixes: ’Neu’ (neurologist), ’Pat’ (patient), ’Ap’ (accompanying person). Note, that in Elsey *et al.* features are not extracted on the basis of the neurologists’ speech, but we decided to include these to investigate the role of the neurologists as well. The profiling characteristics in Table 1 are designed for manual extraction; they are largely depending on human language understanding such as the sophistication required to detect the subtle difference in language use between a ’dunno’-style answer indicating that the patient cannot remember, versus indicating something the patient is unsure about. It was decided that for this initial study, no attempt would be made to introduce higher level language understanding. Instead, shallow features have been proposed in the hope of capturing information correlated with the Elsey features.

Many of the linguistic features are *textual* in nature and a common natural language (NLP) approach known as Bag-of-Words (BoW) model [13] is used for the extraction. This technique underpins many search engines (like Google) and is supported by numerous NLP packages (e.g. NLTK [14]). The BoW ignores the order of words, punctuations, commonly used words in English, and trims the verbs to their original stems. The remaining linguistic features can be thought of as *conceptual* features defined on the interaction turn-level. To extract those, an automatic way of splitting the conversation into questions and answers is needed. This is a challenging task, however, in this data we have a relatively structured interaction between patient and neurologist with new topics in the conversations being almost exclusively initiated by the latter. This means we can use a far simpler approach that relies on detecting particular words or phrases in a turn. This facilitates the extraction of features that relies on knowledge of turn boundaries.

[1] Elsey *et al.* call these ’features’, however to avoid confusion, we will refer to them here as ’profiling characteristics’ and use the term ’features’ as is conventional in the speech community.

Table 1: *List of profiling characteristics and corresponding proposed features; see text for further details. (a:acoustic feature, b:syntactic feature, c:semantic feature, and d:visual feature)*

Diagnostic profiling characteristic (Elsey *et al.*)	Proposed, automatic feature(s)
F1) Accompanying person (role of)	number of turns (**1.APsNoOfTurns**^{a}, **2.PatNoOfTurns**^{a}); average length of turn ([sec]) (**3.APsAVTurnLength**^{a}, **4.PatAVTurnLength**^{a}); average unique words in a turn (**5.APsAVUniqueWords**^{b}, **6.PatAVUniqueWords**^{b})
F2) Responding to neurologists' questions about memory problems	patient answered "me" (**7.PatMeForWhoConcerns**^{c})
F3) Patient recall of recent memory failure	number of empty words (**8.PatFailureExampleEmptyWords**^{c}); average length of pauses (**9.PatFailureExampleAVPauses**^{a}); used all the time (**10.PatFailureExampleAllTime**^{c})
F4) Responding to compound questions	patient replies 'dunno for the expectation question (**11.PatDunnoForExcpectations**^{c}); how many times 'dunno' in combination with turning to AP (**12.PatAVNoOfDunno**^{d}); average number of shaking head (**13.PatAVNoOfShakesHead**^{d}); average number of filler words (**14.PatAVFillers**^{c}); average number of empty words (**15.PatAVEmptyWords**^{c}); average number of low-frequency words (**16.PatAVAllWords**^{b})
F5) Inability to answer	average number of repeated questions (**17.AVNoOfRepeatedQuestions**^{c})
F6) Patient's elaborations and length of turn	patients average unique words in a turn (**6.PatAVUniqueWords**^{b}, **4.PatAVTurnLength**^{a})
Features not in Elsey *et al.* but relating to neurologists role	number of turns (**18.NeuNoOfTurns**^{a}); length of turns([sec]) (**19.NeuAVTurnLength**^{a}); average number of unique words (**20.NeuAVUniqueWords**^{b}); average number of topics discussed (**21.AVNoOfTopicsChanged**^{c}); average length of pauses by patient (**22.PatAVPauses**^{a})

A few of the features are related to the length of pauses, and length of turns which are easily extracted from ASR output.

3.3. Automatic speech recogniser

For the purposes of this study, manual speaker turn segmentation was used. Very short segments (less than 0.5 second) as well as overlapping segments were removed. The final data compromised of around 10 hours spontaneous speech from a total of 105 speakers (note that the same doctors were present in multiple interviews), with 7820 utterances (mean utterance length of 4.6 sec). Using the Kaldi toolkit [15] and following the standard WSJCam0 recipe, a baseline speaker-independent ASR (SI_WSJCam0) was trained and then testing with the full dataset (Dem39). This initial baseline model gave a WER of 95.1% (see table 2). The next model was obtained by MAP adapting SI_WSJCam0 onto the Dem39 (SI_WSJCam0+MAP_Dem39). The adaptation process did not seem truly successful (with 81.1% WER), therefore a model was trained from scratch (SI_Dem39). To train and test on this ASR, the held out approach was used with around 20% of data for testing and 80%for training. The average WER was 40.6 %. This relatively high WER reflects the very challenging nature of the data even when the overlapping sections have been excluded.

3.4. Classification

In this study, the focus was on differentiating between ND and FMD groups, so a binary machine learning classifier was used. However, as mentioned before, there is no single classifier that can perform best all the time and depending on data we can choose several classifiers (using a standard validation approach) to find the best classifier with the highest accuracy. Scikit-learn [16] is a Python library with a wide range of machine learning classifiers. From this library, 5 classifiers were chosen: Support Vector Machine (SVM) with linear kernel, Random Forest, Adaptive Boost (Adaboost), Perceptron, and linear classification via Stochastic Gradient Descent (SGD). All accuracies are averaged over cross validation tests on the trained classifiers using the leave-one-out technique resulting in a total of 39 test datapoints.

Table 2: *Speech recognition results.*

Model	WER [%]
SI_WSJCam0	95.1
SI_WSJCam0+MAP_Dem39	81.1
SI_Dem39	40.6

4. Results

4.1. Automatic CA-style feature extraction

To validate how good the features proposed in Table 1 are, an initial experiment was carried out where the features were extracted from the original, verbatim transcripts produced with the Elsey data. This would correspond to having a *perfect* ASR module. The five classifiers described above were trained and the results are presented in Figure 2. The accuracy of the classifiers range from 82 % to 95 %, with an average of 87 %; the best

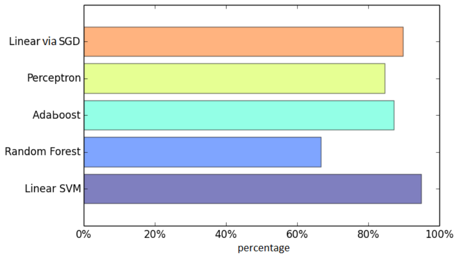

Figure 2: *Classifier accuracies using all 22 extracted features.*

classifier was the SVM with an average cross-validation accuracy of 95 %. This is comparable to the human classification achieved with recent Elsey *et al.* (in preparation) results where they demonstrated that CA-based profiles enabled the correct diagnosis of ND/FMD in 10 out 10 cases with one conversation analyst and 9 out of 10 with another conversation analyst.

For the subsequent experiments, results for the Linear SVM are reported. This system includes all the 22 features described in Table 1. However, two of those features are based on the video stream and are therefore unavailable for the fully-automatic audio-based processing used in the following experiments, and a smaller feature set with 20 audio-only features was therefore used. Repeating the above experiments resulted in a 92 % classification accuracy (referred to as AutoCA in the following, with the full feature set named AutoCA(visual)).

4.2. Effect of adding automatic speech recognition

As the WER results in Table 2 indicate, the data is challenging and the discrepancies between the manual and the automatic transcripts are huge. Table 3 shows how this affects the classification accuracy when the features are extracted from the ASR transcripts rather than from the verbatim transcripts. Each row in the table represents an increase in ’automatisation’ from the fully manual method in the first row to the fully automatic system in row 4. The last two systems apply feature selection and is further discussed in section 4.3.

Comparing the obtained accuracies for AutoCA to ASR+AutoCA shows that introducing the ASR reduces the classification accuracy to 79 %. However, the different types of features (acoustic, syntactic and semantic) are very differently affected as can be seen in Figure 3. For the AutoCA system, the semantic features were the most informative (accuracy of 85 %) but as expected, the noisy ASR transcripts means that these features become less meaningful (dropping to 67 %). Similarly, the acoustic features drop in accuracy from 77 % to 67 %. In contrast, the syntactic features become more significant and their accuracy increases from 67 % to 82 %.

4.3. Effect of feature selection

The last two rows of Table 3 show the effect of doing feature selection and only using the top 10 features by Recursive Feature Elimination (RFE) [16]. For both the AutoCA and the ASR+AutoCA system, the feature selection boosts result. For ASR+AutoCA the accuracy increases from 79 % back up to 90 %, which is close to the AutoCA value of 95 %. The rankings of the top 10 informative features for AutoCA and the ASR+AutoCA are listed in Table 4. Most of these features are shared by both systems (in bold) but with a different ranking, and they are mostly related to the patients, however, the neurologist’s wordings was also important for the classification. This indicates the role of the conversational partner when communicating with people with dementia, which has been reported by other authors e.g. [17, 18].

Table 3: *Linear SVM classifier’s accuracy for the transcripts.*

System	Transcript	CA-feature	Accuracy
Elsey *et al.* in prep.	manual	manual(6)	95 %
AutoCA(visual)	manual	automatic(22)	95 %
AutoCA	manual	automatic(20)	92 %
ASR+AutoCA	ASR	automatic(20)	79 %
AutoCA	manual	auto+feat.sel.(10)	95 %
ASR+AutoCA	ASR	auto+feat.sel.(10)	90 %

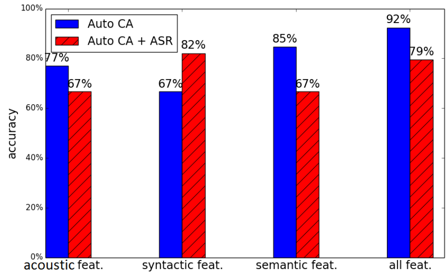

Figure 3: *Linear SVM classifier’s accuracy for AutoCA and AutoCA+ASR using different types of features*

Table 4: *Top 10 features for the Auto CA and ASR+AutoCA.*

Rank	AutoCA	ASR+AutoCA
1	**PatAVAllWords**	APsAVUniqueWords
2	**PatFailureExampleAVPauses**	PatNoOfTurns
3	**NeuAVUniqueWords**	**PatAVAllWords**
4	**APsNoOfTurns**	**PatAVUniqueWords**
5	PatDunnoForExcpectations	**PatFailureExampleAVPauses**
6	**AVNoOfTopicsChanged**	**NeuAVUniqueWords**
7	PatMeForWhoConcerns	**APsAVTurnLength**
8	**APsAVTurnLength**	**APsNoOfTurns**
9	**PatAVUniqueWords**	PatAVTurnLength
10	PatAVPauses	**AVNoOfTopicsChanged**

5. Conclusions

We have presented an automatic system for extracting CA inspired features from conversations between neurologists and patients with suspected dementia. We have demonstrated that the proposed features enable a very high ND/FMD discrimination accuracy despite being shallow representations emulating human language understanding. When coupling this with ASR and feature selection algorithms we demonstrate that the accuracy only drops from 95 % to 90 % despite the very challenging data and WER around 40 %. This is highly encouraging for the methods proposed; future work will concentrate on expanding the feature set and improving on the speech recognition by introducing e.g., automatic diarization.

6. References

[1] S. Bell, K. Harkness, J. M. Dickson, and D. Blackburn, "A diagnosis for 55: what is the cost of government initiatives in dementia case finding." *Age and Ageing.*, vol. 44, pp. 344–345, 2015.

[2] A. J. Larner, "Impact of the National Dementia Strategy in a neurology-led memory clinic: 5-year data." *Clin Med.*, vol. 14, p. 216, 2014.

[3] C. Elsey, P. Drew, D. Jones, D. Blackburn, S. Wakefield, K. Harkness, A. Venneri, and M. Reuber, "Towards diagnostic conversational profiles of patients presenting with dementia or functional memory disorders to memory clinics," *Patient Education and Counseling.*, vol. 98, pp. 1071–1077, 2015.

[4] D. Jones, P. Drew, C. Elsey, D. Blackburn, S. Wakefield, K. Harkness, and M. Reuber, "Conversational assessment in memory clinic encounters: interactional profiling for differentiating dementia from functional memory disorders," *Aging & Mental Health.*, vol. 7863, pp. 1–10, 2015.

[5] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care," *Interspeech.*, pp. 1781–1784, 2005.

[6] R. J. Moore, "Automated Transcription and Conversation Analysis," *Research on Language and Social Interaction.*, vol. 48, no. 3, pp. 253–270, 2015.

[7] K. López-de Ipiña, J.-B. Alonso, C. M. Travieso, J. Solé-Casals, H. Egiraun, M. Faundez-Zanuy, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, and U. Martinez de Lizardui, "On the selection of non-invasive methods based on speech analysis oriented to automatic Alzheimer disease diagnosis." *Sensors.*, vol. 13, pp. 6730–45, 2013.

[8] K. López-de Ipiña, J. Solé-Casals, H. Eguiraun, J. Alonso, C. Travieso, A. Ezeiza, N. Barroso, M. Ecay-Torres, P. Martinez-Lage, and B. Beitia, "Feature selection for spontaneous speech analysis to aid in Alzheimer's disease diagnosis: A fractal dimension approach," *Computer Speech & Language.*, vol. 30, pp. 43–60, 2015.

[9] W. Jarrold, B. Peintner, D. Wilkins, D. Vergryi, C. Richey, M. L. Gorno-Tempini, and J. Ogar, "Aided diagnosis of dementia type through computer-based analysis of spontaneous speech," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pp. 27–37, 2014.

[10] C. Thomas, V. Keselj, Cercone, K. Rockwood, and E. Asp, "Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech," *Proceedings of the IEEE International Conference on Mechatronics & Automation.*, pp. 1569–1574, 2005.

[11] K. C. Fraser, J. M. A, and F. Rudzicz, "Linguistic Features Identify Alzheimer s Disease in Narrative Speech," *J Alzheimers Dis.*, vol. 49, pp. 407–22, 2015.

[12] M. Yancheva, K. Fraser, and F. Rudzicz, "Using linguistic features longitudinally to predict clinical scores for Alzheimer's disease and related dementias," *6th Workshop on Speech and Language Processing for Assistive Technologies.*, 2015.

[13] G. Salton, *Introduction to modern information retrieval.* New York: McGraw-Hill, 1983.

[14] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python.* Sebastopol, CA: OReilly Media Inc., 2009.

[15] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.* IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[16] F. Pedregosa and G. Varoquaux, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research.*, vol. 12, pp. 2825–2830, 2011.

[17] H. E. Hamilton, *Conversations with an Alzheimers patient: An interactional sociolinguistic study.* Cambridge: Cambridge University Press, 1994.

[18] L. Perkins, A. Whitworth, and R. Lesser, "Conversing in dementia: A conversation analytic approach," *Journal of Neurolinguistics.*, vol. 11, pp. 33–53, 1998.