SAMS: Data and Text Mining for Early Detection of Alzheimer’s Disease

November, 2016
Dr Christopher Bull
Aim of talk

- What is SAMS
- Data Capture
  - Problems and solutions to acquiring this type of text/data
- NLP
  - Tools used
    - Existing
    - Bespoke
- Reflections
Who am I?

Dr Christopher Bull

c.bull@lancaster.ac.uk
@ChrisBull88

• 2011 – PhD
• 2014 – SAMS (PDRA)
• 2016 – Mobile Age (PDRA)

• Software Engineering
• Education/Pedagogy
• Digital Health Technologies

[Insert dashing photo here]
SAMS Overview
• National Dementia Strategy (2009): early (‘timely’) diagnosis

• Only about 50% of people with dementia currently receive a diagnosis

• Diagnosis is often late - moderate or severe stages
What is Alzheimer’s Disease?

Alzheimer’s is the most common cause of dementia (estimated 60%-80% of cases)

- Dementia “describes symptoms that occur when the brain is affected by certain diseases or conditions”

Symptoms include:

- memory loss
- difficulties with:
  - thinking
  - problem-solving
  - language

Ultimately fatal

Source: Alzheimer’s Society
Goal:
Explore Technology-dependent proxy markers of Alzheimer’s Disease

Aims:
• Non intrusive capture of computer use
• Mine the data for trends and patterns
• Infer longitudinal changes in cognitive health
<table>
<thead>
<tr>
<th>Name</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professor Pete Sawyer</td>
<td>School of Computing and Communications, Lancaster University</td>
</tr>
<tr>
<td>Dr Paul Rayson</td>
<td>School of Computing and Communications, Lancaster University</td>
</tr>
<tr>
<td>Dr Christopher Bull</td>
<td>School of Computing and Communications, Lancaster University</td>
</tr>
<tr>
<td>Professor Alistair Sutcliffe</td>
<td>School of Computing and Communications, Lancaster University</td>
</tr>
<tr>
<td>Professor Alistair Burns</td>
<td>National Clinical Director for Dementia in England, Institute of Brain,</td>
</tr>
<tr>
<td></td>
<td>Behaviour and Mental Health, University of Manchester</td>
</tr>
<tr>
<td>Dr Iracema Leroi</td>
<td>Institute of Brain, Behaviour and Mental Health, University of Manchester</td>
</tr>
<tr>
<td>Gemma Stringer</td>
<td>Institute of Brain, Behaviour and Mental Health, University of Manchester</td>
</tr>
<tr>
<td>Dr Samuel Couth</td>
<td>Institute of Brain, Behaviour and Mental Health, University of Manchester</td>
</tr>
<tr>
<td>Professor John Keane</td>
<td>School of Computer Science, University of Manchester</td>
</tr>
<tr>
<td>Dr Ann Gledson</td>
<td>School of Computer Science, University of Manchester</td>
</tr>
<tr>
<td>Professor Clive Ballard</td>
<td>Wolfson Centre for Age-Related Diseases, King’s College London</td>
</tr>
</tbody>
</table>
Data Flows

Raw data → Data mining → Cognitive indicators

Sources of non-text data

Sources of text

Raw text → Text mining

Bayesian reasoning over evidence

Inferred cognitive health

SAMS user
Current Status

- Project funding ended September 2016
- On-going analysis
My Role in SAMS

...and Data Collection
My Role

• Data capture software
  – Software Design/implementation
    • SAMS Manager
    • Browser extensions
  – Maintenance (obviously)

• Text Mining
  – Text extraction (reconstruction)
  – Reusing existing NLP pipeline (Wmatrix; UCREL)
  – Implementing extensions to pipeline for specific heuristics

• General Project Support (Team & Participants)
• Consider challenges
Challenges

• Volatility of participant computers
  – Unexpected updates
  – Varying shutdown procedures
  – Various software setups (anti-virus etc.)
• Weak performing computers (and not monopolise valuable resources)
  – Again, various hardware/software setups
• Ethical challenges
  – Privacy/Security
• Novel monitoring approaches
• Internet Explorer *sigh*
• Win 10 roll-out mid project
Abstract Architecture (Data Collection)

Collecting context, not just raw data

Desktop/Application Monitor Processes

Encrypt Logs

Secure SAMS Server

Manager Process

Browser Extensions
Desktop/Application Monitor Processes

<table>
<thead>
<tr>
<th>Desktop/App Monitor</th>
</tr>
</thead>
<tbody>
<tr>
<td>▶️ C# input event listeners</td>
</tr>
<tr>
<td>▶️ Variety of Mouse, keyboard.</td>
</tr>
<tr>
<td>▶️ Windows Automation API: UI Automation (UIA)</td>
</tr>
<tr>
<td>▶️ Observe UI elements (and properties) a user interacts with.</td>
</tr>
<tr>
<td>▶️ Provides context behind events.</td>
</tr>
</tbody>
</table>

*Work of Dr Ann Gledson, Mancs*
Browser Extensions

Browser Extension

Webpage black/whitelist (e.g. no https:// unless predefined)

JS DOM parsing (text fields and interactive elements)

JS event listeners & context identifier (Click, Mouse-Move, Focus etc.)

Log message caching (volatile)

Encryption

Write log files
Browser Monitoring - Challenges

- Context to events
- Constantly changing or dynamic DOM
Manager/Uploader

• Process management
• Server communication
• Remote updating
• Log message caching and encryption
Manager (2)

SAMS Study: Pause/Resume Monitoring

We would like to confirm that you are Chris:

- Yes, I am
- No, I am not

Pause monitoring
Project Support

- Participant Status Checker
  - For clinical & Tech teams
  - +Android app

- Phone support
  - Clinical Team
  - Participants

- Participant visits (Installs)
## Existing Study(s)

### Nun Study:
- Measures obtained from autobiographies written over a 60-year span (age 22 to 83).

<table>
<thead>
<tr>
<th></th>
<th>No dementia</th>
<th>Dementia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical complexity</td>
<td>-mean 4.78</td>
<td>-mean 3.86</td>
</tr>
<tr>
<td></td>
<td>-declined .04 units per year</td>
<td>-declined .03 units per year.</td>
</tr>
<tr>
<td>Idea density</td>
<td>-mean 5.35</td>
<td>-mean 4.34</td>
</tr>
<tr>
<td></td>
<td>propositions per 10 words</td>
<td>propositions per 10 words</td>
</tr>
<tr>
<td></td>
<td>- declined .03 units per year</td>
<td>-declined .02 units per year.</td>
</tr>
</tbody>
</table>

Language Decline Across the Life Span: Findings From the Nun Study

Susan Kemper  
University of Kansas

Lydia H. Greiner  
University of Kentucky

Janet G. Marquis, Katherine Prenovost, and Tracy L. Mitzner  
University of Kansas
Propositional Idea Density (P-density)

• “Idea density [...] is the number of expressed propositions divided by the number of words. In terms of semantics, idea density is a measure of the extent to which the speaker is making assertions (or asking questions) rather than just referring to entities”
  – “Automatic measurement of propositional idea density from part-of-speech tagging” (Brown et al, 2008)

• Existing Implementation
  – CPIDR (Computerized Propositional Idea Density Rater)
  – (pronounced “spider”)
  – only tool to automate this*

* At time of starting SAMS
Kusari (Toolchain manager)

“Toolchain and data dependency manager for use with conventional NLP toolchains”

Dr Steve Wattam

https://delta.lancs.ac.uk/Steve/kusari
https://delta.lancs.ac.uk/Steve/kusari-links
# Toolchain

<table>
<thead>
<tr>
<th>Spelling Variation</th>
<th>VARD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="ucrel.lancs.ac.uk/vard/">ucrel.lancs.ac.uk/vard/</a></td>
</tr>
<tr>
<td></td>
<td>Java</td>
</tr>
<tr>
<td>Part Of Speech Tagger</td>
<td>CLAWS</td>
</tr>
<tr>
<td></td>
<td><a href="ucrel.lancs.ac.uk/claws/">ucrel.lancs.ac.uk/claws/</a></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Semantic Tagger</td>
<td>USAS</td>
</tr>
<tr>
<td></td>
<td><a href="ucrel.lancs.ac.uk/usas/">ucrel.lancs.ac.uk/usas/</a></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>Frequency Lists</td>
<td>Tmatrix</td>
</tr>
<tr>
<td></td>
<td><a href="ucrel.lancs.ac.uk/wmatrix/">ucrel.lancs.ac.uk/wmatrix/</a></td>
</tr>
<tr>
<td></td>
<td>C</td>
</tr>
<tr>
<td>SAMS software</td>
<td>SNOWCAT</td>
</tr>
<tr>
<td></td>
<td><a href="delta.lancs.ac.uk/SAMS/SNOWCAT">delta.lancs.ac.uk/SAMS/SNOWCAT</a></td>
</tr>
<tr>
<td></td>
<td>Java</td>
</tr>
</tbody>
</table>
Sams aNalysis of Output from Wmatrix for the Cognitive Assessment of Text

- Input
  - Tmatrix (FQLs)
  - USAS (Sem)

- Output
  - CSV of metrics
SNOWCAT: Sample Output (1/2)

- Total Words (MWE), 26278
- Total Words, 27787
- Vocabulary size (MWE), 3533
- Vocabulary size, 3444
- **Type:Token (ratio; MWE),** 0.134
- **Type:Token (ratio),** 0.124
- **Type:Token (normalised ratio), 0.403**
- Words occurring once (MWE), 1842
- Adjective (total; MWE), 1288
- Adjective (ratio; MWE), 0.049
- Noun (total; MWE), 4280
- Noun (ratio; MWE), 0.163
- ...

Lancaster University

SAMS
SNOWCAT: Sample Output (2/2)

- Pronoun (total; MWE), 2672
- Pronoun (ratio; MWE), 0.102
- Verb (total; MWE), 6135
- Verb (ratio; MWE), 0.233
- Content words (total; MWE), 13757
- Content words (ratio; MWE), 0.524
- Filler words (total; MWE), 183
- Filler words (ratio; MWE), 0.007
- Noun:Verb (ratio; MWE), 0.698
- Mean Length of Utterance, 27.653
- VARD Variant (total), 69
- VARD Variant (ratio), 0.003
- Propositional Idea Density, 0.565
Early (unpublished) Results

- Validate P-Density (comparison to CPIIDR tool)
- Uses novelist study to explore usefulness of SNOWCAT metrics
- [Show spreadsheet of early (unpublished) results]
Charts

[Charts removed from this public copy of the presentation, due to the data being pre-publication]
What’s next?

- Continue NLP analysis
- Correlate Data and Text Mining analyses
- ...SAMS 2.0
Lessons Learnt

• Ethical process
  – Affects fundamental design decisions

• Complexity of data collection outside of “lab setting”

• Validating other studies/claims important
Thank you

November, 2016
Dr Christopher Bull

http://ucrel.lancs.ac.uk/sams/
c.bull@lancaster.ac.uk
@ChrisBull88
• Combining data mining and text mining for detection of early stage dementia: the SAMS framework.

• From Click to Cognition: Detecting cognitive decline through daily computer use.

• Dementia and Social Sustainability: Challenges for Software Engineering.

• Discovering affect-laden requirements to achieve system acceptance.