Natural Language Processing of UK Annual Report Narratives

Paul Rayson
Mahmoud El-Haj
Session outline

• What is NLP?
• How we have applied NLP methods to the financial domain in the CFIE project
• Demo of our systems
  – Wmatrix-import
  – Wmatrix
Natural Language Processing

• ‘getting computers to understand language’ … ‘a branch of computer science dealing with analysing, understanding and generating natural language’
NLP and Corpus Linguistics @ Lancaster

• Analyzing digital personas behind criminal tactics to provide investigators with clues to the identity of the individual or group hiding behind one or more personas
  – Child sex offenders masquerading as young people to gain their victims’ trust
  – Radicalization of youth in online forums through persuasive messaging
  – Online dating sites that gain trust and the exploit users for financial gain
  – Analysis of cyber bullying and online trolling
  – Use textual analysis to predict a persona’s key attributes (e.g., age and gender) even where the author is obfuscating the language

• Assist in early diagnosis of dementia and Alzheimer’s disease
  – Longitudinal changes in linguistic features (e.g., grammatical complexity and idea density) are associated with degenerative memory conditions
Corporate Financial Information Environment (CFIE)

• ESRC- and ICAEW-funded project
  – Martin Walker, Manchester Business School
  – Steven Young, Lancaster University Management School
  – Paul Rayson, Lancaster University School of Computing & Communications
  – Mahmoud El-Haj, Lancaster University School of Computing & Communications
  – Vasiliki Athanasakou, London School of Economics
  – Thomas Schleicher, Manchester Business School

• Project seeks to analyse UK financial narratives, their association with financial statement information, and their informativeness for investors

• Automated, large sample analysis of UK annual report narratives represents a cornerstone of the project
• Develop software for general use by academics
• http://ucrel.lancs.ac.uk/cfie/
Analysing UK Annual Report Narratives

Harvest reports

Collect documents manually or use crawler to automate collection

Clean and parse text

e.g., use mark-up tags to parse sections, remove exhibits and pictures; clean prior to analysis, etc.

Analyze text

Word counts based on wordlists

Measuring constructs using off-the-shelf software

Machine learning (training software to recognise patterns)

Text mining to identify patterns

Depends on file type (*.pdf, *.rtf, *.txt, etc.)
US Filings: 10-K Annual Form

Each 10-K contains 4 parts and 15 items

- **PART I**
  - **ITEM 1.** Description of Business
  - **ITEM 2.** Description of Properties
  - **ITEM 3.** Legal Proceedings
  - **ITEM 4.** Mine Safety Disclosures

- **PART II**
  - **ITEM 5.** Market for Registrant’s Common Equity….
  - **ITEM 6.** Selected Financial Data
  - **ITEM 7.** Management’s Discussion and Analysis….
  - **ITEM 8.** Financial Statements and Supplementary Data
  - **ITEM 9.** Changes in and Disagreements ….

- **PART III**
  - **ITEM 10.** Directors, Executive Officers and Corporate Governance
  - **ITEM 11.** Executive Compensation
  - **ITEM 12.** Security Ownership of Certain Beneficial Owners….
  - **ITEM 13.** Certain Relationships and Related Transactions….
  - **ITEM 14.** Principal Accounting Fees and Services

- **PART IV**
  - **ITEM 15.** Exhibits, Financial Statement Schedules….
c.f. UK Annual Reports

- Free Style (no standard structure)
- Use of Images, Text
- PDF Format
- Content and structure varies across firms.
- Management have more discretion over what, where, and how much information on topics such as risk, strategy, performance, etc. is reported.

- This makes the extraction and analysis task more challenging; but it provides research opportunities.
• Use contents page to extract text by section from digital pdf
## Contents

<table>
<thead>
<tr>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>56</td>
<td>Consolidated income statement</td>
</tr>
<tr>
<td>57</td>
<td>Consolidated balance sheet</td>
</tr>
<tr>
<td>58</td>
<td>Consolidated cash flow statement</td>
</tr>
<tr>
<td>59</td>
<td>Reconciliation of net cash flow to movements in net debt</td>
</tr>
<tr>
<td>60</td>
<td>Consolidated statement of recognised income and expense</td>
</tr>
<tr>
<td>61</td>
<td>Consolidated statement of Changes in equity</td>
</tr>
<tr>
<td>101</td>
<td>Company cash flow statement</td>
</tr>
<tr>
<td>102</td>
<td>Company statement of recognised income and expense</td>
</tr>
<tr>
<td>103</td>
<td>Company statement of changes in equity</td>
</tr>
<tr>
<td>112</td>
<td>Notes to the Company financial statements</td>
</tr>
<tr>
<td>113</td>
<td>Five year record</td>
</tr>
<tr>
<td>114</td>
<td>Shareholder analysis</td>
</tr>
<tr>
<td>115</td>
<td>Financial calendar</td>
</tr>
<tr>
<td>116</td>
<td>Investor information</td>
</tr>
<tr>
<td>100</td>
<td>Company balance sheet</td>
</tr>
<tr>
<td>99</td>
<td>Independent auditors' report – Company</td>
</tr>
<tr>
<td>54</td>
<td>Statement of directors' responsibilities</td>
</tr>
<tr>
<td>42</td>
<td>Directors' remuneration report</td>
</tr>
<tr>
<td>36</td>
<td>Corporate governance report</td>
</tr>
<tr>
<td>33</td>
<td>Directors' report</td>
</tr>
<tr>
<td>32</td>
<td>Board of directors and company secretary</td>
</tr>
<tr>
<td>04</td>
<td>Chief executive's report</td>
</tr>
<tr>
<td>02</td>
<td>Chairman's statement</td>
</tr>
<tr>
<td>08</td>
<td>Performance review</td>
</tr>
<tr>
<td>26</td>
<td>Corporate social responsibility report</td>
</tr>
<tr>
<td>IFC</td>
<td>Financial and operating highlights</td>
</tr>
</tbody>
</table>

## Financial highlights:

- Continuing profit before taxation* up 25.3% at £50.1m (2006/07: £40.0m)
- Profit for the period £34.5m (2006/07: loss £22.5m)
- Adjusted EPS¹ 20.9% ahead at 7.88p (2006/07: 6.52p)
- Strong cash management continues to drive low net debt² at £200.2m (2006/07: £174.2m)
- Robust balance sheet – continuation of share buy-back programme in 2008/09
- Proposed dividend up 5.9% at 4.50p (2006/07: 4.25p)
What's in our report

Features

Sir Ian Gibson
We are committed to making food shopping fresh, friendly and affordable.
2 Page 2

Dalton Philips
A clear strategy is in place that is delivering our objectives.
4 Page 4

Richard Pennycook
Our strong financial performance positions us well for sustainable long term growth.
6 Page 6

Johanna Waterous
A strong performance culture, long term shareholder value and competitive positioning remain key principles.
8 Page 8

Also see...

Corporate responsibility review 2011/12

Annual review 2011/12

Investor relations website

www.morrisons.co.uk/corporate

Directors' report and business review

Introduction
2 Chairman's statement

Business and strategy review
4 Chief Executive's business and strategy review
8 Group Finance Director's financial review
12 Our strategic objectives

Performance review
26 Key performance indicators
28 Risks and uncertainties
30 Corporate responsibility
33 Our people

Governance
36 Board of Directors and Management Board
40 Corporate governance report
46 Directors' remuneration report
56 General information
59 Statement of Directors' responsibilities

Financial statements
60 Group financial statements
60 Independent auditor's report
61 Consolidated statement of comprehensive income
62 Consolidated balance sheet
63 Consolidated cash flow statement
64 Consolidated statement of changes in equity
65 Group accounting policies
70 Notes to the Group financial statements

95 Company financial statements
95 Company balance sheet
96 Company accounting policies
99 Notes to the Company financial statements

Investor information
108 Five year summary of results
109 Supplementary information
110 Investor relations and financial calendar
<table>
<thead>
<tr>
<th>Page</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>02</td>
<td>Who we are and what we do</td>
<td>42</td>
</tr>
<tr>
<td>06</td>
<td>24 hours in the life of Arriva</td>
<td>46</td>
</tr>
<tr>
<td>08</td>
<td>Our growth story</td>
<td>47</td>
</tr>
<tr>
<td>10</td>
<td>Our markets</td>
<td>48</td>
</tr>
<tr>
<td>12</td>
<td>Chairman’s statement</td>
<td>52</td>
</tr>
<tr>
<td>14</td>
<td>Chief executive’s review</td>
<td>56</td>
</tr>
<tr>
<td>22</td>
<td>Financial review</td>
<td>82</td>
</tr>
<tr>
<td>26</td>
<td>Corporate responsibility</td>
<td>83</td>
</tr>
<tr>
<td>32</td>
<td>Board of directors</td>
<td>90</td>
</tr>
<tr>
<td>34</td>
<td>Directors’ report</td>
<td>91</td>
</tr>
<tr>
<td>37</td>
<td>Directors’ remuneration report</td>
<td>92</td>
</tr>
</tbody>
</table>

Corporate governance
Statement of directors’ responsibilities
Independent auditors’ report on the group financial statements
Financial statements
Accounting policies
Notes to the accounts
Five-year financial summary
Parent company financial statements
Statement of directors’ responsibilities on the parent company financial statements
Independent auditors’ report on the parent company financial statements
Financial calendar, registered office and advisors
Steps in extraction process:
- Detect contents page
- Parse contents page
- Detect page numbering to determine section start/end
- Add headers as bookmarks to pdf
- Extract text for each section

Analyse extracted text by section and for entire document
In addition to performing text extraction, the tool provides a range of text analysis options:

- Readability metrics
- Word counts using pre-determined lists (e.g., forward looking, uncertainty, tone, etc.)
- Word counts based on user-defined wordlists
- Comparison with reference corpus (word, parts of speech and semantic level)
- Concordance and collocates
- Upload and analyse user-defined text file
• Overview (wordlists, readability metrics) and interface with Wmatrix
• Uploading one or more annual reports and generating output
• Uploading and analysing with a user-defined key word list
• Uploading and analysing a user-defined text file
• Examples of further analysis in Wmatrix:
  – Cloud for chairman’s statement vs. standard reference corpus: word level
  – Cloud for chairman’s statement vs. to standard reference corpus: semantic level
  – Cloud for chairman’s statement vs. chairman’s statement corpus: word level
  – Cloud for chairman’s statement vs. chairman’s statement corpus: semantic level
  – Concordance/collocation
Case study: Strategic content

- Measure strategic content of annual reports by automatic analysis of words and phrases in financial commentary and business strategy sections
- Prior UK studies: manual, 300 reports
- 60,000 sections in 10,000 reports
- 1156 key words and phrases
- Clear evidence of doubling of strategic content since 2002
Future developments

• Choice of different readability metrics
• Text re-use metric to detect boilerplating and incremental changes
• Reference corpora for:
  – Full Annual Report and Accounts
  – Chairman’s Statement
  – Highlights
  – Corporate Governance
  – Directors’ Remuneration
• Readability scores for all UK annual reports for the period 2003-2013
• Further methods for representing key themes (e.g., strategy, risk, CSR and governance, etc.) by section