CORPORATE FINANCIAL INFORMATION ENVIRONMENT (CFIE)

MAHMOUD EL-HAJ
SCC, LANCASTER UNIVERSITY
TEAM AND FUNDING

Prof Martin Walker  
MBS

Prof Steve Young  
LUMS

Dr Paul Rayson  
SCC

Dr Mahmoud El-Haj  
SCC

Dr Vasiliki Athanasakou  
LSE

Dr Thomas Schleicher  
MBS
Primary ways that firms communicate with capital market participants.

Together with information from:

- analysts,
- financial journalists,
- rating agencies and
- other market commentators that are external to the firm

combine to form the Corporate Financial Information Environment (CFIE)
study the causes and consequences of corporate disclosure and financial reporting outcomes.
aim to uncover the determinants of financial reporting quality
and the factors that influence the quality of information disclosed to investors beyond the financial statements.
we aim to scale up the application of current readability metrics and improve their granularity.

to improve on previous work, we need to apply the metrics to individual sections of firms' annual reports.

a necessary prerequisite for our work is to automatically determine the structure of these reports.
- 1,500 searchable financial annual reports
- of around 200 of the largest UK firms listed on the LSE
- with an average of 7 annual reports for each firm
- between the years 2003 and 2012.
CFIE ANALYSIS PIPELINE

Annual Reports (PDFs)

PDF-to-Text

Extract Headers

Extract Sections

NLP(SEM)

Readability Metrics

Display Results (Web)
APPLY NLP METHODS USED IN PRIOR US STUDIES TO UK ANNUAL REPORTS?

UK VS. US FILINGS
US FILINGS

- US companies must submit:
  1. 10-K: Annual
  2. 10-Q: Quarterly
  3. 8-K: Special Events (between 10-K and 10-Q)
  4. Annual Report
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….
UK ANNUAL REPORTS

- Free style (no standard structure)
- Use of images, text, hyperlinks, …etc.
- 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.
UK ANNUAL REPORTS SAMPLE
EXTRACTION PROCESS

WHAT ARE WE LOOKING TO EXTRACT?

C  F  I  E
We are looking to extract the following headers and their narratives for further processing:

1. Chairman’s statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report
5. Directors Report and Business Review
6. Directors Responsibilities Statement
7. Directors Report
8. Financial Review
9. Key Performance Indicator
10. Operational Review
11. Highlights
Contents

Spirax Sarco at a glance ........................................ 6
Chairman’s statement ........................................ 8
Business review ................................................. 10
   Market overview ........................................... 10
   Performance review ....................................... 15
Board of Directors ............................................. 28
Directors’ report ................................................ 31
   Corporate governance ..................................... 34
   Corporate social responsibility ......................... 38
   The Directors’ remuneration report ................... 42
   Statement of Directors’ responsibilities .............. 50

Chairman’s overview
by Sir Stuart Rose

HOW?

Not consistent across ARs

Doesn’t always refer to the correct page
1) detecting the contents-page
2) parsing the detected contents-page and extracting the headers
3) detecting page numbering
4) adding the extracted headers to the annual report PDFs as bookmarks
5) using the added bookmarks to extract the narrative sections under each heading

The processes run on searchable (text-based) PDFs; we will consider using OCR techniques to process non-searchable (scanned) PDFs in a later stage.
I) DETECTING THE CONTENTS PAGE

- created a list of gold-standard section names extracted manually from a random sample of 50 annual reports
- matched each page in the annual report against the gold-standard list
- selected the page with the highest matching score as the potential contents page
- the score was calculated by an increment of 1 for each match.
- To improve the matching process and avoid false positives, we match the gold–standard keywords against lines of text that follow a contents-page-like style (e.g. section name followed by page number, such as Chairman’s Statement 13).
2) PARISING THE CONTENTS PAGE

- We automatically parsed the detected contents page to extract section names and their associated pages.
- Matched each line of text in the potential contents page against a regular expression command that will extract any line starting or ending with a number between 1 and the number of pages of the annual report.
- We differentiate between dates and actual page numbers to avoid extracting incorrect section headers.
- However, lines containing text such as an address (e.g., 77 London Road) might still be confused.
- We tackled this problem by matching the list of extracted headers against a list of gold-standard header synonyms.
- To tackle the problem of broken headers we concatenating sentences that end or begin with prepositions such as ‘of’, ‘in’ …etc.
- The algorithm also concatenates sentences ending with singular or plural possessives, symbolic and textual connectors (e.g. ‘and’, ‘or’, ‘&’…etc), and sentences ending with hyphenations.
The page numbers appearing on the contents page do not usually match with the actual page numbers in the PDF files.

Created a simple page detection tool that crawls through a dynamic number of three consecutive pages with the aim of extracting a pattern of sequential numbers with an increment of 1 (e.g. 31, 32, 33).

Running this process we got an accuracy rate of 94%.

Manual examination of the remaining 6% revealed the following reasons for non-detection: 1) encoding, 2) formatting and 3) design.
4) ADDING HEADERS AS BOOKMARKS

- Using the headers and their correct page numbers we implemented a tool to insert the extracted contents page headers as bookmarks (hyperlinks) to sample PDFs.
- This process helped in extracting narratives associated with each header for further processing.
5) EXTRACTING HEADERS’ NARRATIVES (PART 1)

- Automatically crawl through the data collection and extract all inserted bookmarks and their associated pages.
- Since UK firms do not follow a standard format when creating annual reports, a long list of synonyms are possible for a single header.
- For example the header “Chairman’s Statement” may also appear as “Chairman's Introduction”, “Chairman's Report” or “Letter to Shareholders”.
- To solve this problem we, semi automatically and by the help of an expert in accounting and finance, created a list of synonyms for each of the 11 generic annual report headers.
- This was done by extracting all headers containing “Chairman”, “Introduction”, “Statement”, “Letter to”…etc from a sample of 250 annual reports of 50 UK firms (the quoted unigrams were selected by the same expert).
- We refined the list by removing redundancies. The accounting expert then manually examined the list and deleted irrelevant or inappropriate headers.
- We used the refined list as gold–standard synonyms to extract all the headers related to each of our generic headers.
To tackle different word–order or additional words included in the headline (e.g. “The Statement of the Chairman”) we used Levenshtein Distance string metric algorithm to measure the difference between two headers.

The Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, substitution) required to change one word into the other.

To work on a sentence level we modified the algorithm to deal with words instead of characters.

All the headers with a Levenshtein distance of up to five were presented to the accounting expert.

1. Chairman’s statement
2. CEO Review
3. Corporate Government Report
4. Directors Remuneration Report
5. Directors Report and Business Review
6. Directors Responsibilities Statement
7. Directors Report
8. Financial Review
9. Key Performance Indicator
10. Operational Review
11. Highlights
For a sample of 250 annual reports we analysed each report and its extracted sections by calculating text readability scores using Flesh and Fog readability measures.

We also counted word frequencies using forward looking, hedging, positive and negative words–lists.
FORWARD LOOKING FREQUENCIES
To ensure quality, we used domain experts to judge the quality of the document structure extraction process.

We took a random sample of 100 previously unseen annual reports that had bookmarks automatically added to them through the extraction process.

The expert human evaluators were presented with an evaluation form and asked to compare the automatically assigned bookmarks to the contents page of the same annual report.

An expert in the accounting and finance domain went through the extracted headers and their narrative sections to judge the quality of the extraction process, the expert also updated the gold–standard list with any new unseen synonyms.

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Year</th>
<th>Number of Headers in PDF</th>
<th>Number of Extracted Headers</th>
<th>Number of Exact Matches</th>
<th>Number of Partial Matches</th>
<th>Number of Wrong Headers</th>
<th>Page Numbers Correct?</th>
<th>If NO, what is the difference between PDF and Report page numbers?</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 ERGO 31AUG04</td>
<td>2004</td>
<td>31</td>
<td>24</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JI GROUP PLC_07</td>
<td>2007</td>
<td>40</td>
<td>39</td>
<td>38</td>
<td>0</td>
<td>1</td>
<td>Yes</td>
<td></td>
<td>Picked a footer</td>
</tr>
<tr>
<td>ACAL PLC_09</td>
<td>2009</td>
<td>24</td>
<td>24</td>
<td>22</td>
<td>2</td>
<td>0</td>
<td>No</td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
The evaluators’ input was used to calculate Recall/Precision and F measure.

The manual evaluation was performed in two separate stages following the same evaluation process.

Stage 1 helped identify the most common errors that led to incorrect extraction and detection of either the contents page and its headers or the annual report’s page numbering.

Stage 2 was performed after fixing errors discovered by the human evaluators.
### EVALUATION: STAGE 1 AND 2

<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th>Stage 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Percent</td>
</tr>
<tr>
<td># of PDFs</td>
<td>105</td>
<td>-</td>
</tr>
<tr>
<td>Headers in PDFs</td>
<td>2473</td>
<td>-</td>
</tr>
<tr>
<td>Extracted Headers</td>
<td>2479</td>
<td>-</td>
</tr>
<tr>
<td>Exact Matches</td>
<td>2101</td>
<td>84.8%</td>
</tr>
<tr>
<td>Partial Matches</td>
<td>189</td>
<td>7.6%</td>
</tr>
<tr>
<td>Wrong Headers</td>
<td>189</td>
<td>7.6%</td>
</tr>
<tr>
<td>Missing Headers</td>
<td>183</td>
<td>7.4%</td>
</tr>
<tr>
<td>Correct Headers</td>
<td>2290</td>
<td>92.6%</td>
</tr>
<tr>
<td>Detected Page number</td>
<td>80</td>
<td>76.2%</td>
</tr>
<tr>
<td>Detected Contents Pages</td>
<td>97</td>
<td>92.4%</td>
</tr>
</tbody>
</table>
An extracted header is considered ‘strictly relevant’ only if it is an exact match of a PDF's header.

The header is considered ‘broadly relevant’ if it is either an exact match or a partial match of a PDFs header.

Results reveal the fixes applied helped increase recall and precision rates by extracting more relevant headers.
THANKS
ANY QUESTIONS

CFIE UREL: http://ucrel.lancs.ac.uk/cfie