Computer-based Analysis of the Strategic Content of UK Annual Report Narratives

Vasiliki Athanasakou Mahmoud El-Haj Paul Rayson Martin Walker Steven Young

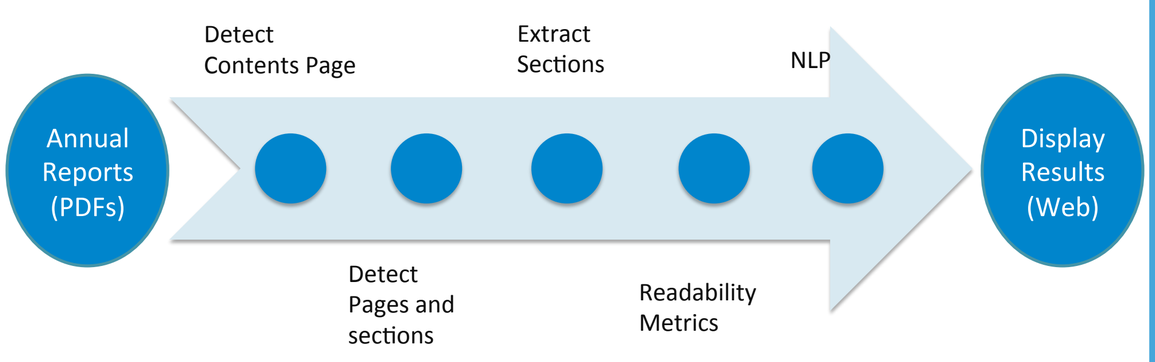
PURPOSE AND MOTIVATION

Study the causes and consequences of corporate disclosure and financial reporting outcomes. The determinants of financial reporting quality and the factors that influence the quality of information disclosed to investors beyond the financial statements are the main focal points of the project. Important to detect document structure due to the unstructured nature of the information in the UK Annual Financial Reports. The UK cases contrasts with the situation in the US where the structure of financial filings is defined and fixed.

DATA COLLECTION

- Annual Reports of UK Firms
- ~10,000 Searchable PDFs
- ~200,000 section headings in aggregate
- Majority of firms listed on the London Stock Exchange
- Sample period is 2002-2013

NLP AND READABILITY TOOLS



OVERVIEW OF STRUCTURING PROCESS

1. Detect the contents page
2. Parse the contents page
3. Use a synonym list to match expected headers list
4. Detect page numbering
5. Add headers as bookmarks
6. Extract text under each header

PARSING THE CONTENTS PAGE

1. Match each line of text against a regular expression (e.g. text – number)
2. Extracted page numbers are between one and max pages in the annual report
3. Differentiate between page numbers, dates and values (e.g. £32)
4. Avoid addresses (e.g., 77 London Road) by matching to the gold-standards
5. Tackle broken headers using algorithm to detect incomplete lines (e.g., detect lines ending with “and”, “or”, “in”, … etc.)

DETECTING PAGE NUMBERING

1. Created a page number detection tool
2. Crawls through a dynamic number of three consecutive pages
3. Aim to extract a patter of numbers with a +1 increment (e.g.3, 4, 5)
4. Algorithm yields accuracy rate of 94%

HEADERS LIST

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

EVALUATION

- Used domain experts to judge the quality
- Took a random sample of 100 unseen annual reports with auto bookmarks
- Human evaluators compared the bookmarks to the contents page
- The evaluators completed a form recording the number of partial/exact matches
- Evaluators added comments to explain their evaluations
- An expert in accounting and finance went through the extracted headers and evaluators’ recorded sections to judge quality and update the gold-standards when necessary
- The evaluators input was used to calculate recall and precision, and F1 scores
- The evaluation was divided into two stages 1 and 2
- Stage 2 was performed after fixing errors discovered by the human evaluators

RESULTS

	Stage 1		Stage 2	
	Count	%	Count	%
# of PDFs	105	-	105	-
Headers in PDFs	2,473	-	2,473	-
Extracted Headers	2,479	-	2,502	-
Exact Matches	2,101	84.80%	2,202	88.01%
Partial Matches	189	7.60%	105	4.20%
Wrong Headers	189	7.60%	195	7.80%
Missing Headers	183	7.40%	166	6.60%
Correct Headers	2,290	92.60%	2,307	93.30%
Detected Page number	80	76.20%	94	89.50%
Detected Contents Pages	97	92.40%	97	92.40%

Corporate Financial Information Environment (CFIE)
Project URL:
http://ucrel.lancs.ac.uk/cfie/