View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Lancaster E-Prints

Detecting Document Structure in a Very Large Corpus of UK Financial Reports

Mahmoud El-Haj*, Paul Rayson*, Steven Young, Martin Walker*****

*School of Computing and Communications, Lancaster University, UK
**Lancaster University Management School, Lancaster University, UK
***Manchester Business School, Manchester University, UK
*{m.el-haj, p.rayson}@lancaster.ac.uk
**s.young@lancaster.ac.uk
***martin.walker@mbs.ac.uk

Abstract

In this paper we present the evaluation of our automatic methods for detecting and extracting document structure in annual financial reports. The work presented is part of the Corporate Financial Information Environment (CFIE) project in which we are using Natural Language Processing (NLP) techniques to study the causes and consequences of corporate disclosure and financial reporting outcomes. We 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. The CFIE consists of the supply of information by firms to investors, and the mediating influences of information intermediaries on the timing, relevance and reliability of information available to investors. It is important to compare and contrast specific elements or sections of each annual financial report across our entire corpus rather than working at the full document level. We show that the values of some metrics e.g. readability will vary across sections, thus improving on previous research research based on full texts.

Keywords: document structure, annual reports, readability

1. Introduction

In previous accounting literature, readability scores have been used to measure the linguistic characteristics of key corporate disclosures, to identify determinants of cross–sectional variation in these characteristics, and to relate these characteristics to disclosure informativeness. Previous research has also calculated word frequencies using forward looking, hedging, positive and negative words–lists. This research has only been possible on a small scale due to the manual nature of the task. We aim to scale up the application of such metrics and improve their granularity. In the research beyond that presented here, we have employed a suite of statistical and rule-based NLP tools for analysing firms' narrative communication practices. We hypothesise that there will be significant variability of such measures. In addition, we need to focus on some report sections while removing others e.g. with only quantitative numerical data. To improve on previous work, we therefore need to apply the metrics to individual sections of firms' annual reports. Hence, a necessary prerequisite for our work is to automatically determine the structure of these reports. Figure 1 shows the complete analysis process constructed for the CFIE project[1] and where the document structure extraction is located in the process.

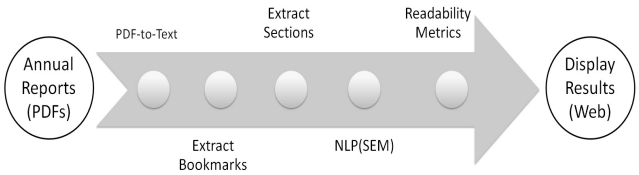

Figure 1: CFIE Analysis Pipeline

2. Related Work

Previous work on detecting document structure was conducted on collections of books, newspapers and scientific articles. Power et al. (2003) recognised the importance of modelling document structure for natural language processing and natural language generation, arguing that it should be distinguished from rhetorical structure. Teufel (2010) provides a book level treatment of the subject aimed at discovering discourse structure in scientific articles.

Competitions to extract structure from documents have been run to stimulate research in this area. For example, Doucet et al. (2009) describes the book structure extraction competition which was run during the 2009 International Conference on Document Analysis and Retrieval (ICDAR). The task was to automatically extract hyperlinked tables of contents from previously digitised books. A modified Levenshtein edit distance measure was used to compare the relative match between automatically extracted titles and a manually built gold–standard.

Recent activities have also extended structure extraction beyond logical document elements to incorporate discourse structures such as the modelling of argumentation, narrative and rhetorical structure in the 'Detecting Structure in Scholarly Discourse' (DSSD2012) workshop[2].

3. Data Collection

The data used in our experiments consisted of 1,500 financial annual reports of around 200 of the largest UK firms listed on the London Stock Exchange, with an average of seven annual reports for each firm between the years

[1] http://ucrel.lancs.ac.uk/cfie/

[2] http://www.nactem.ac.uk/dssd/index.php

2003 and 2012. The annual reports vary in respect to their style and number of pages. In contrast to the USA, stock exchange-listed firms in UK do not present their financial information and accompanying narratives in a standardised format when creating annual reports. Instead, UK firms have much more discretion regarding the structure and content of the annual report. Added to this is the problem of nomenclature: no standardised naming convention exists for different sections in UK annual reports so that even firms adopting the same underlying structure and content may use different terminology to describe the same section(s). Finally, whereas financial filings made by firms in the USA are presented in plain text format, UK firms' annual reports are published as PDFs. The combination of these factors makes identifying document structure much more problematic for UK firms compared to their counterparts in the USA.

4. Structure Extraction Process

To detect and extract the structure of our annual reports, we applied the following five main processes: 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, and 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 but image-based PDFs are in the minority in our dataset with only 10% of the larger dataset.

4.1. Detecting the Contents Page

An annual report contents page includes information about the main sections of the report and its associated page numbers. Information in the contents page helped us detect the structure of the annual report. However, detecting the contents page was not a straightforward task. We created a list of gold–standard section names extracted manually from a random sample of 50 annual reports. We matched each page in the annual report against the list of section names, then we 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*).

4.2. Parsing the Contents Page

We automatically parsed the detected contents page to extract section names and their associated pages. To do this we 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 Bothwell Road) might still be confused. We tackled this problem by matching the list of extracted headers against a list of gold–standard header synonyms (see Section 4.5. below). To tackle the problem of broken headers (i.e., headers appearing on two lines or more), we implemented an algorithm to detect broken sentences and fix them by 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.

4.3. Detecting Page Numbering

The page numbers appearing on the contents page do not usually match with the actual page numbers in the PDF files. For example, page 4 in the annual report could refer to page 6 in the PDF file, which may lead to incorrect extraction. We addressed this problem by creating a 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 on the sample yielded 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.

4.4. Adding Headers as Bookmarks

Using the headers and their correct page numbers from Sections 4.1. and 4.3. 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 (see Section 4.5. below).

4.5. Extracting Headers' Narratives

We implemented an automatic extraction tool to crawl through the data collection and, for each PDF file, 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 generic annual report headers (see the list below). 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 (e.g. all headers about the "Chairman's Statement"). To overcome the problem of different word–order or additional words included in the headline (e.g. "The Statement of the Chairman") we used *Levenshtein Distance* string metric algorithm (Levenshtein, 1966) to measure the difference between two headers. The Levenshtein distance between two words is the minimum number of single-character edits (insertion, deletion, sub-

stitution) 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.

We used the above process to create gold–standard synonym lists for the following 11 generic headers that we wished to extract for further analysis:

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

Having detected and extracted headers (or their gold–standard synonyms) and their sections, we then extract the sections' narratives using iText[3], an open source library to manipulate and crate PDF documents (Lowagie, 2010), to apply our text analysis metrics, which include readability measurement and counting word frequencies using financial domain hand–crafted word lists. We are aware of the limitations of counting using word lists, but wish to replicate earlier studies on a larger scale and will then move on to address this issue in further work.

5. Analysis and Readability Measures

For a sample of 250 annual reports (see Section 4.5.) we analysed each report and its extracted sections by calculating text readability scores using Flesh and Fog readability measures (Gunning, 1968), (Kincaid et al., 1975). We also counted word frequencies using forward looking, hedging, positive and negative words–lists. Figure 2 shows the Flesch readability measure of the annual reports (AR) sample compared to the 'Chairman's Statement' (CMS) section of each of those reports. One conclusion we could make is that the readability of the Chairman's statement generally reflects the readability of the whole annual report. However, some reports do not follow this trend so previous research considering readability over complete reports will need to be revisited.

[3]http://itextpdf.com/api

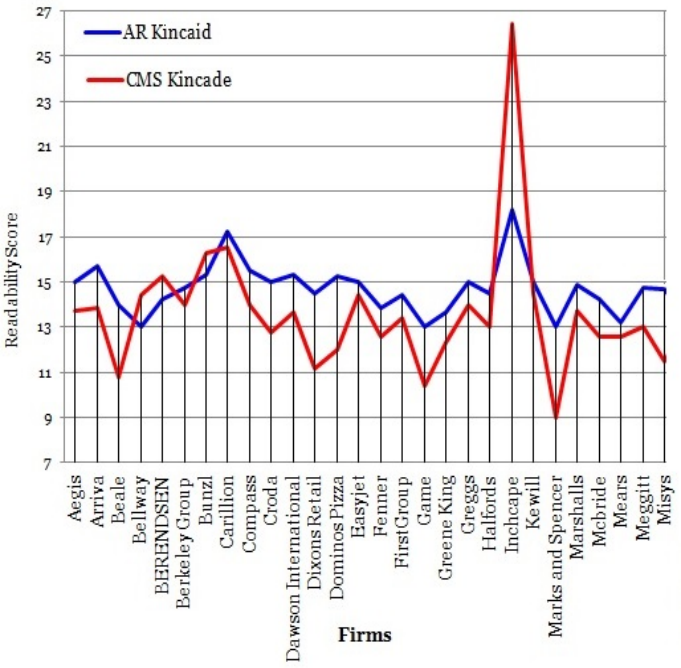

Figure 2: Readability: Annual Report vs Chairman's Section

6. Evaluation

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 process described in Section 4. 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. The evaluators reported the number of matching headers, exact or partial, and whether the automatic detection of the contents page and the page numbering were correct. The evaluators also added comments and notes to explain their evaluations.

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 of any of the 11 headers (see Section 4.5.). The evaluators' input was used to calculate recall, precision and F1 scores, the evaluation metrics are defined as follows:

$$Precision = \frac{relevant\_headers\_retrieved}{all\_retrieved\_headers}$$

$$Recall = \frac{relevant\_headers\_retrieved}{all\_relevant\_headers}$$

$$F1\_Score = 2 * \frac{precision * recall}{precision + recall}$$

7. Results

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

in Section 6. 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. Table 1 illustrates the evaluation results of Stage 1 and Stage 2, respectively.

	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	84.8%	2,202	88.01%
Partial Matches	189	7.6%	105	4.20%
Wrong Headers	189	7.6%	195	7.8%
Missing Headers	183	7.4%	166	6.6%
Correct Headers	2,290	92.6%	2,307	93.3%
Detected Page number	80	76.2%	94	89.5%
Detected Contents Pages	97	92.4%	97	92.4%

Table 1: Evaluation: Stage 1 and 2

Stage 2 results in Table 1 shows an improvement in the number of extracted headers when compared to the results of Stage1, with more exact matched and fewer partially matched headers. The fixes applied helped reduce the number of missing headers and improved the algorithm's ability to correctly detect annual report page numbering, with a success rate close to 90%.

Table 2 shows the Recall, Precision and F1-Score results of Stage 1 and Stage 2 evaluations. 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 PDF's header. Results reveal the fixes applied helped increase recall and precision rates by extracting more relevant headers.

	Stage 1	Stage 2
Strict Recall	0.8496	0.8904
Strict Precision	0.8475	0.8801
Strict F-1 Score	0.8485	**0.8852**
Broad Recall	0.9260	0.9329
Broad Precision	0.9238	0.9221
Broad F-1 Score	0.9249	**0.9274**

Table 2: Recall/Precision/F1 Scores: Stage 1 and 2

8. Conclusion

In this paper, we have described a process to reliably extract document structure from PDF documents representing annual financial reports. This forms a necessary preprocessing step prior to our NLP pipeline which carries out further analysis and calculation of a variety of metrics such as readability, hedging and forward-looking language on a section-by-section basis as well as on full reports. Better understanding of the structure of these documents will, we hope, provide a much more fine-grained analysis of the narratives contained in these annual financial reports. It will allow us to compare and contrast specific sections over time and across different companies and sectors and this will in turn allow us to better understand the corporate financial information environment. In future work, we will apply these techniques on a very large scale to over 10,000 reports spread over 10 years and consider the variation in reporting quality and, with the benefit of hindsight, how it correlates to subsequently reported financial performance.

Acknowledgement

The project is funded by the Economic and Social Research Council (ESRC) (Ref. ES/J012394/1) and the Institute of Chartered Accountants in England and Wales (ICAEW).

9. References

Doucet, Antoine, Kazai, Gabriella, Dresevic, Bodin, Uzelac, Aleksandar, Radakovic, Bogdan, and Todic, Nikola. (2009). Icdar 2009 book structure extraction competition. In *Proceedings of the Tenth International Conference on Document Analysis and Recognition (ICDAR'2009)*, pages 1408–1412, Barcelona, Spain, July.

Gunning, R. (1968). *The Technique of Clear Writing*. McGraw-Hill.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index fog count and flesch reading ease formula) for navy enlisted personnel. In *Research Branch Rep*, pages 8–75, Memphis, TN.

Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Technical Report 8.

Lowagie, B. (2010). *iText in Action*. Covers iText 5. Manning Publications Company.

Power, Richard, Scott, Donia, and Bouayad-Agha, Nadjet. (2003). Document structure. *Comput. Linguist.*, 29(2):211–260, June.

Teufel, Simone. (2010). *The structure of scientific articles: Applications to Citation Indexing and Summarization*. CSLI Studies in Computational Linguistics. Center for the Study of Language and Information, Stanford, California.