MultiLing 2019: Financial Narrative Summarisation

Mahmoud El-Haj
School of Computing and Communications
Lancaster University
United Kingdom
m.el-haj@lancaster.ac.uk

Abstract

The Financial Narrative Summarisation task at MultiLing 2019 aims to demonstrate the value and challenges of applying automatic text summarisation to financial text written in English, usually referred to as financial narrative disclosures. The task dataset has been extracted from UK annual reports published in PDF file format. The participants were asked to provide structured summaries, based on real-world, publicly available financial annual reports of UK firms by extracting information from different key sections. Participants were asked to generate summaries that reflects the analysis and assessment of the financial trend of the business over the past year, as provided by annual reports. The evaluation of the summaries was performed using AutoSummENG and Rouge automatic metrics. This paper focuses mainly on the data creation process.

1 Introduction

Firms and businesses worldwide use a number of different methods to communicate with their shareholders and investors and to report to the financial markets. These include annual financial reports, quarterly reports, preliminary earnings announcements, conference calls and press releases (El-Haj et al., 2018a).

For the financial narrative summarisation task we focus on annual reports produced by UK firms listed on the London Stock Exchange (LSE). In the UK and elsewhere, annual report structure is much less rigid than those produced in the US, and companies produce glossy brochures with a much looser structure, and this makes automatic summarisation of narratives in UK annual reports a challenging task since the structure of those documents needs to be extracted first in order to summarise the narrative sections of the annual reports. This can happen by detecting narrative sections that usually includes the management disclosures rather than the financial statements of the annual reports.

2 Related Work

The volume of available information is increasing sharply and therefore the study of NLP methods that automatically summarise content has grown rapidly into a major research area. At the conceptual level, text summarisation is the process of distilling content of a single document or a set of related documents down to the most important events presented in the correct sequence. Automatic text summarisation is therefore the process of producing a condensed version of a text using computerised methods. The aim is for the summary to convey the key contributions of the original text. Automated text summarisation therefore involves identifying key sentences. The process of defining key sentences is highly dependent on the summarisation method used.

The ongoing MultiLing series[1] tailored tasks towards multilingual single and multi-document summarisation aimed towards pushing the state of the art in automatic text summarisation and this year Multiling is introducing the first Financial Narrative Summarisation task focused towards English UK annual reports (Li et al., 2013; El-Haj et al., 2011; Elhadad et al., 2013; Giannakopoulos et al., 2011).

Cardinaels et al. (2018) is the only Accounting and Finance study of which we are aware that uses statistical and heuristic summarisers to generate summaries of financial disclosures. Results reveal that automatic algorithm-based summaries of earnings releases are generally less positively biased than management summaries, and that in-

[1] http://multiling.iit.demokritos.gr/

Proceedings of the Multiling 2019 Workshop, co-located with the RANLP 2019 conference, pages 6–10
Varna, Bulgaria, September 6, 2019.
http://doi.org/10.26615/978-954-452-058-8_002

vestors who receive an earnings release accompanied by an automatic summary arrive at more conservative valuation judgements.

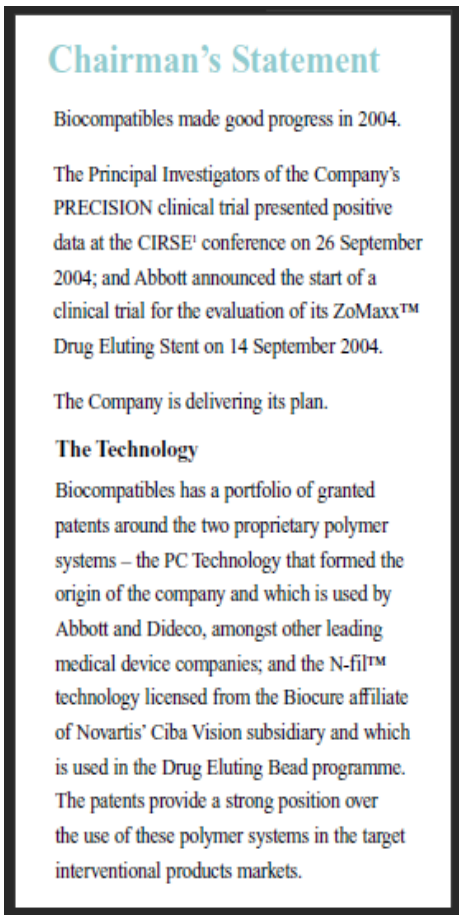
de Oliveira et al. (2002) created a summarisation system that uses lexical cohesion[2] to summarise financial news collected from Reuters' Website[3].

3 Data Description

Before we indulge into describing the summaries dataset we start by a short introduction of what an annual report is. Firms in the UK and worldwide produce an annual document called an 'annual report' which provides a comprehensive reporting on a company's activities throughout the preceding year. Annual reports are intended to give shareholders and other interested parties information about the company's activities and financial performance. They may be considered as grey literature. It was not until legislation was enacted after the stock market crash in 1929 that the annual report became a regular component of corporate financial reporting. Typically, an annual report will contain the following (El-Haj et al., 2019b):

- Financial Highlights
- Letter to the Shareholders
- Narrative Text, Graphics and Photos
- Management's Discussion and Analysis
- Financial Statements
- Notes to Financial Statements
- Auditor's Report
- Summary Financial Data
- Corporate Information

Annual reports are usually long documents spanning between 60 and up to 300 pages. As the reports are provided in PDF file format, extracting strcture is a challening task. The work by (El-Haj et al., 2018b, 2019b) used the UK annual report's table of contents to retrieve the textual content (narratives) for each section listed in the table of contents. Section headings presented in the table of contents are used to partition retrieved content into the audited financial statements component of the report and the "front-end" narratives component, with the latter sub-classified further into a set of generic report elements including the Chairman's Statement, CEO Review, the Governance Statement, the Remuneration Report, and report's Highlights. Figure 1 shows a narrative example extracted from the Chairman's Statement Section in front-end of an annual report.



Chairman's Statement

Biocompatibles made good progress in 2004.

The Principal Investigators of the Company's PRECISION clinical trial presented positive data at the CIRSE[1] conference on 26 September 2004; and Abbott announced the start of a clinical trial for the evaluation of its ZoMaxx™ Drug Eluting Stent on 14 September 2004.

The Company is delivering its plan.

The Technology

Biocompatibles has a portfolio of granted patents around the two proprietary polymer systems – the PC Technology that formed the origin of the company and which is used by Abbott and Dideco, amongst other leading medical device companies; and the N-fil™ technology licensed from the Biocure affiliate of Novartis' Ciba Vision subsidiary and which is used in the Drug Eluting Bead programme. The patents provide a strong position over the use of these polymer systems in the target interventional products markets.

Figure 1: Front-End narratives example - Chairman's Statement

To detect the structure of UK annual reports we used the CFIE-FRSE software to detect structure for around 4000 UK annual reports for firms listed on LSE covering the period between 2002 and 2017 (El-Haj et al., 2014, 2019a,c; El Haj et al., 2018). CFIE-FRSE stands for Corporate Financial Information Environment (CFIE) -Final Re-

[2]Lexical cohesion refers to the way related words are chosen to link elements of a text. There are two forms: repetition and collocation. Repetition uses the same word, or synonyms, antonyms, etc. For example, "Which dress are you going to wear?" - "I will wear my green frock" uses the synonyms "dress" and "frock" for lexical cohesion. Collocation uses related words that typically go together or tend to repeat the same meaning. An example is the phrase "once upon a time".

[3]http://www.reuters.co.uk

port Structure Extractor (FRSE). The tool is available as a desktop application, which is freely available on GitHub[4]. The tool detects the structure of annual reports by detecting the key sections, their start and end pages in addition to the narrative contents.

Using CFIE-FRSE we divided the annual reports' full text into *training*, *testing* and *validation*. We also provide the sections extracted using CFIE-FRSE and we indicate which sections are the "narrative" sections, thus containing the textual contents of the annual reports (see Section 5 below for more details on how we define narrative sections).

For the creation of the financial narrative summarisation dataset we used a number of 3,863 annual reports. We randomly split the dataset into training (c75%), testing and validation (c25%). Table 1 shows the dataset details. We provided the participants with the training and validation datasets including the full text of each annual report along with the extracted sections and gold-standard summaries. At a later stage the participants were given the testing data. On average there are at least 2 gold-standard summaries for each annual report. We do not provide the PDF annual reports and instead we provide the full text as plain text file.

Table 1: Dataset

Data Type	Training	Testing	Validation	Total
Report full text	3,000	500	363	3,863
Report sections	60,794	12,089	9,247	82,130
Gold summaries	6,787	1,151	878	8,816

4 Task Description

In this task We introduce a new summarisation task which we call '**Sturece-based Summarisation**'. In this task the summary requires extraction from different key sections found in the annual reports. Those sections are usually referred to as "narrative sections" or "front-end" sections and they usually contain textual information and reviews by the firm's management and board of directors. Sections containing financial statements in terms of tables and numbers are usually referred to as "back-end" sections and are not supposed to be part of the narrative summaries.

For the purpose of this task we ask the participants to produce one summary for each annual report. The summary length should not exceed **1000** words. We advise that the summary is generated/extracted based on the narrative sections (see Section 5, therefore the participating summarisers need to be trained to detect narrative sections before creating the summaries.

Figure 2 shows the structure of the Financial Narrative Dataset. At the beginning of the shared task we provided the participants with two directories "training" and "validation" each containing the full text of the annual reports (*_full_text), the extracted sections (*_sections) and the gold standard summaries (*_gold_standards).

5 Data Sample

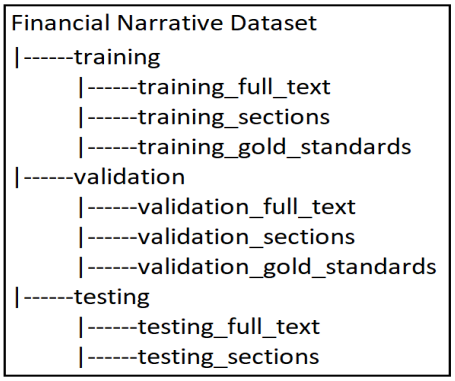


Figure 2: Dataset Structure

The data is provided in plain text file format in a directory structure as in Figure 2. Each annual report has a unique ID and it is used across in order to link annual reports' full text to their sections and gold-standards. For example: The *training* directory contains a file called **17.txt** where 17 is a unique ID and can be used to locate this report's sections in the *training_sections* directory, as shown in the files **17_896317_3.txt** and **17_896317_4.txt**. Also the same ID can be used to find this report's gold standard summaries as in the *training_gold_standards* as in the files **17_896311_8.txt** and **17_896313_1.txt**.

For the files in the **_sections* and **_gold_standards* each file name is made of the following: **reportID_sectionID_sectionType.txt** as in **17_896317_4.txt**.

Section type can be used to identify narrative sections, those with any *sectionType* but zero, as follows:

- 1 Chairman's statement

[4] https://github.com/drelhaj/CFIE-FRSE

- 2 Chief Executive Officer (CEO) review
- 3 Governance statement
- 4 Remuneration report
- 5 Business review
- 6 Financial review
- 7 Operating review
- 8 Highlights
- 9 Auditors report
- 10 Risk management
- 11 Chairman's governance introduction
- 12 Corporate Social Responsibility (CSR) disclosures

Sections with *sectionType*=0 are considered to be non-narratives and are not expected to appear in the summary. Example: **17_896315_0.txt**. To make the task challenging we did not provide section types in the testing data as that is expected to be the participants task where they are expected to define which sections are narrative sections before summarising the report.

The data is available for free for research pruposes.[5]

6 Challenges

This is a challenging task considering a) the size of each annual reports and b) the lack of standardisation in UK annual reports. These challenges shed light on the complexity of financial narratives in general, along with the fact that more robust and up to date machine learning and NLP techniques are required to facilitate the automatic extraction and analysis of financial narratives.

7 Conclusion and Future Work

This paper introduces the first financial narrative summarisation dataset at the First MultiLing Financial Narrative Summarisation Task, held at MultiLing 2019 Summarisation workshop at RANLP 2019 in Varna, Bulgaria . It shows the need and as well as the challenges of summarising long and unstructured UK annual reports. For the future work we will provide a baseline summariser reporting AutoSummENG and Rouge automatic metrics.

[5] http://multiling.iit.demokritos.gr/pages/view/1648/task-financial-narrative-summarization

References

Eddy Cardinaels, Stephan Hollander, and Brian J White. 2018. Automatic summaries of earnings releases: Attributes and effects on investorsâ judgments. *Available at SSRN 2904384*.

Mahmoud El-Haj, Paulo Alves, Paul Rayson, Martin Walker, and Steven Young. 2019a. Retrieving, classifying and analysing narrative commentary in unstructured (glossy) annual reports published as pdf files. *Accounting and Business Research*, pages 1–29.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2011. University of essex at the tac 2011 multilingual summarisation pilot.

Mahmoud El-Haj, Paul Rayson, Paulo Alves, Carlos Herrero-Zorita, and Steven Young. 2019b. Multilingual financial narrative processing: Analysing annual reports in english, spanish and portuguese. *Multilingual Text Analysis: Challenges, Models, And Approaches*, page 441.

Mahmoud El-Haj, Paul Rayson, Paulo Alves, Carlos Herrero-Zorita, and Steven Young. 2019c. Multilingual financial narrative processing: Analysing annual reports in english, spanish and portuguese. *Multilingual Text Analysis: Challenges, Models, And Approaches*, page 441.

Mahmoud El-Haj, Paul Rayson, and Andrew Moore. 2018a. The first financial narrative processing workshop (fnp 2018). *LREC 2018*.

Mahmoud El-Haj, Paul Rayson, Steven Young, and Martin Walker. 2014. Detecting document structure in a very large corpus of uk financial reports.

Mahmoud El-Haj, Paul Edward Rayson, Paulo Alves, and Steven Eric Young. 2018b. Towards a multilingual financial narrative processing system. *LREC 2018*.

Mahmoud El Haj, Paul Edward Rayson, Paulo Alves, and Steven Eric Young. 2018. Towards a multilingual financial narrative processing system. *LREC 2018*.

Michael Elhadad, Sabino Miranda-Jiménez, Josef Steinberger, and George Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 2: Czech, hebrew and spanish. *MultiLing 2013*, page 13.

George Giannakopoulos, Mahmoud El-Haj, Benoit Favre, Marina Litvak, Josef Steinberger, and Vasudeva Varma. 2011. Tac 2011 multiling pilot overview.

Lei Li, Corina Forascu, Mahmoud El-Haj, and George Giannakopoulos. 2013. Multi-document multilingual summarization corpus preparation, part 1: Arabic, english, greek, chinese, romanian. In *Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization*, pages 1–12.

Paulo Cesar Fernandes de Oliveira, Khurshid Ahmad, and Lee Gillam. 2002. A financial news summarization system based on lexical cohesion. In *Proceedings of the International Conference on Terminology and Knowledge Engineering, Nancy, France*.