FNS 2020 shared Task Description

Mahmoud El-Haj
Lanacster University, UK

This document describes the Financial Narrative Summarisation Shared Task at the **FNP-FNS 2020 Workshop**.

1 What are financial narratives

Companies produce a variety of reports containing both narrative and numerical information at various times during their financial year, including annual financial reports. This creates a vast financial information environment which can be impossible to keep track. This shows the vital need for automatic summarisers in order to reduce the time and effort of both the shareholders and investors.

2 Data Description

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 challenge task.

For the FNS 2020 Shared task we use around 4,000 UK annual reports for firms listed on LSE covering the period between 2002 and 2017 [El-Haj et al., 2014, El-Haj et al., 2019].

We divided the annual reports' full text into *training*, *testing* and *validation* sets providing both the full text of each annual report along with the gold-standard summaries.

In total there are 3,863 annual reports divided into training, testing and validation sets. Table 2 shows the dataset details.

Data Type	Training	Validation	Testing
Report full text	3,000	363	500
Gold summaries	9,873	1,250	1,673

Table 1: FNS 2020 Shared Task Dataset

3   Data Availability

For the shared task we first provide the training and validation sets, this include the full text of each annual report along with the gold-standard summaries. On average there are at least 3 gold-standard summaries for each annual report with some reports containing up to 7 gold-standard summaries.

The testing dataset will be provided at a later stage. We will not provide the gold-standard summaries for the testing data.

4   Important Dates

- December 1st, 2019: Registration opens.

- February 17th, 2020: Release of training set.

- March 23rd, 2020: Release of test set.

- April 6th, 2020: Registration deadline.

- June 1st, Results Submission deadline.

- September 1st, 2020: Shared Task Papers Due.

- December 12th, 2020: Workshop day.

5   Important URLs

- Shared Task URL: http://wp.lancs.ac.uk/cfie/fns2020/

- Workshop URL: http://wp.lancs.ac.uk/cfie/fnp2020/

- Participation Form: http://bit.ly/34xWCCp

6   Task Description

For the purpose of this task each team should 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, therefore, [optional], your summariser may need to be trained to detect narrative sections before creating the summaries.

7 Data Sample

```
Financial Narrative Dataset
|------training
        |------annual_reports
        |------gold_summaries
|------validation
        |------annual_reports
        |------gold_summaries
|------testing
        |------annual_reports
```
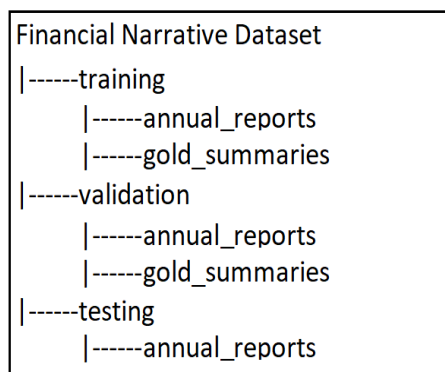
Figure 1: Dataset Structure

Figure 1 shows the structure of the Financial Narrative Summarisation dataset. At the beginning of the shared task we will provide you with two directories "training" and "validation" sets. Each will contain the full text of the annual reports (*_annual_reports) and the gold standard summaries (*_gold_summaries).

The data will be provided in plain text file format in a directory structure as in Figure 1. Each annual report will have a unique ID and it is used across in order to link annual reports' full text to their gold-standard summaries.

For example: The *training/annual_reports* directory contains a file called **19.txt** where 19 is a unique ID and can be used to locate this report's gold standard summaries as in the *training_gold_summaries* as in the files **19_1.txt** to **19_3.txt**.

8 Output

- Each team should produce a no more than 1000 words summary for each annual report in the testing set.

- One summary for each report.

- Each summary should be named following the pattern **ID_summary.txt**. Example: 25082_summary.txt.

- All output should be in UTF-8 file format.

- All output summaries should be compressed following the pattern **TeamName_Summaries.tar.gz**.

References

[El-Haj et al., 2019] El-Haj, M., Rayson, P., Alves, P., Herrero-Zorita, C., and Young, S. (2019). Multilingual financial narrative processing: Analysing annual reports in english, spanish and portuguese.

[El-Haj et al., 2014] El-Haj, M., Rayson, P., Young, S., and Walker, M. (2014). Detecting document structure in a very large corpus of uk financial reports.