# DiMSum: Distributed and Multilingual Summarization of Financial Narratives

Neelesh K Shukla, Amit Vaid, Raghu Katikeri, Sangeeth Keeriyadath, Msp Raja

{nshukla, avaid, rkatikeri, skeeriyadath, smaddila1}@statestreet.com

State Street Corporation, India

Presented at:

The 4th Financial Narrative Processing Workshop (FNP) at LREC, 2022

# FNS 2022 Task

- The objective of the task was to generate not more than 1000 words summaries for the annual financial reports.

- Challenges:
  - Not very rigid structure
  - long reports with 80 pages on average
  - noisy text extracted from PDF
  - variety of information like text, table, financial statements etc. which also deemed repetitive.

- Summary evaluation: ROUGE 2 F1

# FNS 2022 Task Dataset

- FNS-2022 dataset contains annual reports produced by UK , Spanish and Greek firms listed on stock exchange market of each of those countries.

| Type | Training | Validation | Testing |
|---|---|---|---|
| Report Full Text | 3000 | 363 | 500 |
| Gold Summaries | 9873 | 1250 | 1673 |

Table 1: English Dataset

| Type | Training | Validation | Testing |
|---|---|---|---|
| Report Full Text | 162 | 50 | 50 |
| Gold Summaries | 324 | 100 | 100 |

Table 2: Spanish Dataset

| Type | Training | Validation | Testing |
|---|---|---|---|
| Report Full Text | 162 | 50 | 50 |
| Gold Summaries | 324 | 100 | 100 |

Table 3: Greek Dataset

# FNS 2022 Task Dataset- Observations

- Texts extracted form the PDF reports had lot of noise:
  - special characters,
  - unexpected spaces,
  - sentence broken into multiple lines etx
- Gold summaries for the English dataset were extracted directly from the reports
- Less overlap was found in Spanish and Greek datasets.
- Almost all of the English training dataset (99.996%) reports were structured with TOCs.
- The Spanish and Greek reports did not have any reliable TOCs or section headers.

# Hypothesis and Validation

Hypothesis 1: Not everything is relevant and there are some aspects which needs to be focused on. The goal of summarization is to identify salient aspects of report.

Hypothesis 2: Information to be summarized is spread in different aspects which have their own importance and contribution towards summary.

We validated both with the given datasets where summary was generated using sections and at the same time summaries contained different sections based on perspective.

# Problem Statements

- Identify sections or area of report from where the summary needs to be generated.
- Find the importance or weights of identified sections or areas
- Quantify respective contributions towards summary of 1000 words

# Identifying Key Narrative Sections- English

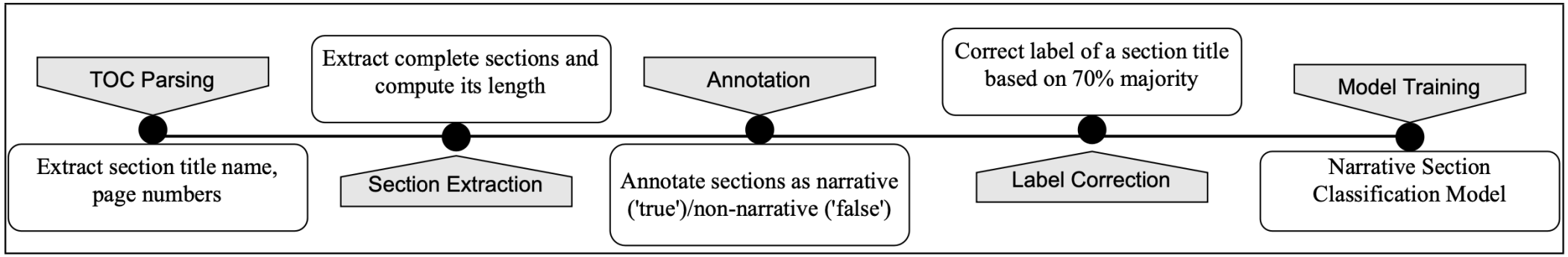



Figure 1: Pipeline for Building Annotated Dataset and Training

- TOC guided
- Binary Classification where section can be narrative ('true') or non-narrative ('false').
- Experimented with many models.
- L2 regularized Logistic Regression with 5-fold CV accuracy 93% and weighted avg f1: .92

| Section Title | #Positive | #Negative |
|---|---|---|
| board of directors | 367 (22%) | **1342 (78%)** |
| chairmans statement | **1729 (72%)** | 668 (28%) |
| chief executives review | **811 (70%)** | 345 (30%) |
| consolidated balance sheet | 152 (13%) | **1012 (87%)** |
| consolidated cash flow statement | 132 (13%) | **872 (87%)** |
| highlights | **713 (75%)** | 240 (25%) |

Table 4: Label Distribution in Annotated Dataset Before Label Correction

# Identifying Key Narrative Sections- English

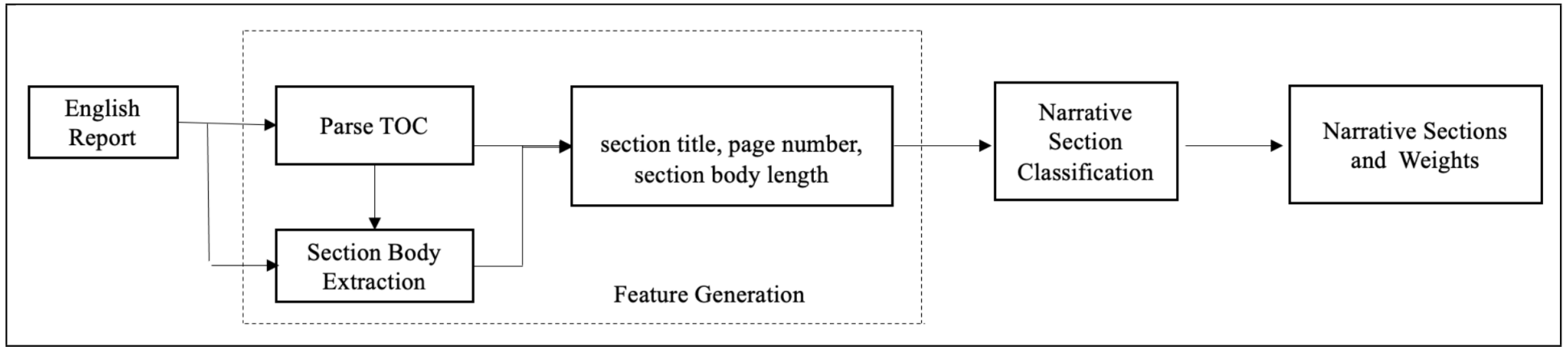



Figure 2: Identifying Key Narrative Sections and Weights in English Reports

Weight of a section can be defined as probability of it being narrative, assigned by the model.

$$W_i : Pr(narrative = true)$$

# Identifying Key Narrative Areas – Spanish/Greek

- We followed TOC guided in English
- Spanish reports – No TOC
- Greek reports – TOCs but can’t be parsed by toc extraction logic which we have written for English
- We also wanted to explore structure (TOC) independent approach
- Focused on identifying a cluster of sentences  as ‘Key Narrative Areas’

# Identifying Key Narrative Areas – Spanish/Greek

**Assumptions**

1. Language Independence
   - The key narratives should be independent of language given all are financial reports.
   - If ’Chairman’s Statement’ is a key narrative in English reports, so ’Declaracio ́n del Presidente’ should be in Spanish.

2. Structure Independence
   - If narratives are not defined as sections, the presence of narrative keywords or key phrases in a sentence indicates it being part of some narrative.

3. Neighbourhoods Assumption
   - If a sentence is part of some narrative, most likely its N neighbouring sentences are also part of the same narrative, defining a set of sentences or paragraph as key narrative area

# Identifying Key Narrative Areas – Spanish/Greek

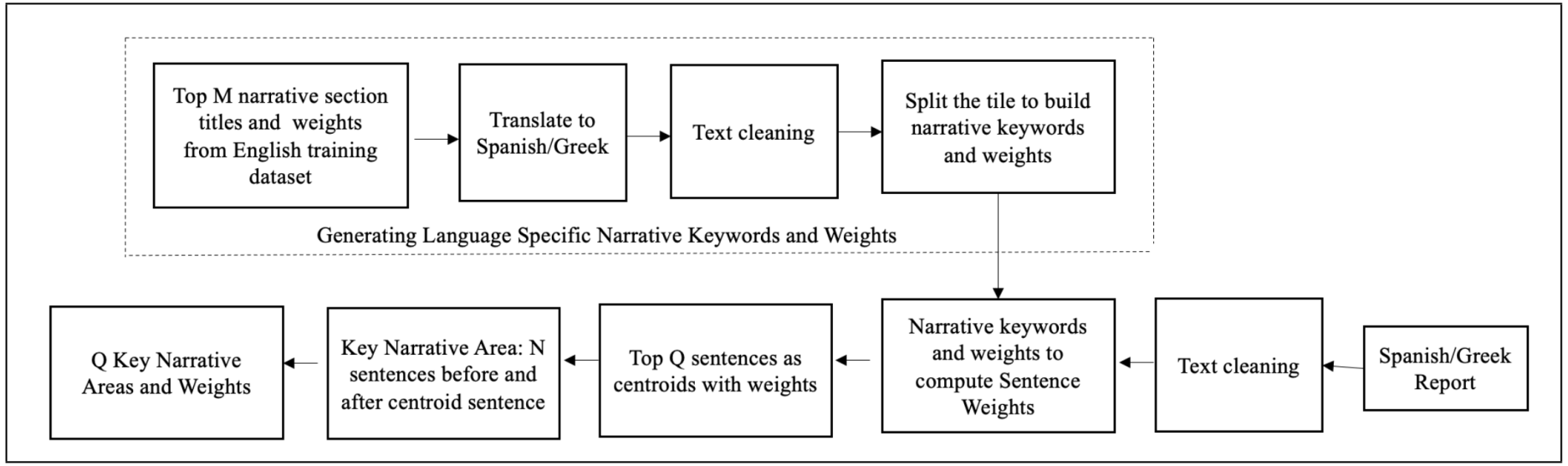



Figure 3: Identifying Key Narrative Areas and Weights in Spanish/Greek Reports

Finetuned Parameters:
M=50, N=20, Q=2

We maintained both raw and processed sentences and summaries were extracted from raw sentences based on position indexes.

# Identifying Key Narrative Areas – Spanish/Greek

- Weight of narrative keywords

$$Wt(w) = \sum Wt(Ns) : w \epsilon Ns$$

where Wt(Ns): Weight of narrative section title Ns.

- Weight of narrative sentence

$$Wt(S) = \sum Wt(w) : w \epsilon S$$

where Wt(w): Weight of narrative keyword w.

Key Narrative Area:

$$[S(i-N), .., Si...., S(i+N)]$$

Weight of Narrative Area:

$$\sum Wt(Sj) : Sj \epsilon [S(i-N), .., Si...., S(i+N)]$$

# Quantify the contributions

- Dividing K= 1000 words/sentences among key narrative areas/sections based on the weights
- Introducing K-Maximal Word Allocation Algorithm

| iter. | section | weight (norm weight) | required #words for summary | #words in section | remaining #words required for summary | remaining #words in section |
|---|---|---|---|---|---|---|
| 1 | section a | 0.90 (0.375) | 375 | 75 | 300 | 0 |
| 1 | section b | 0.90 (0.375) | 375 | 500 | 0 | 125 |
| 1 | section c | 0.60 (0.25) | 250 | 300 | 0 | 50 |

Iteration 1: Required: 1000, Allocated: 700, Remaining Required: 300, Available in Sections: 175

| iter. | section | weight (norm weight) | required #words for summary | #words in section | remaining #words required for summary | remaining #words in section |
|---|---|---|---|---|---|---|
| 2 | section b | 0.90 (0.60) | 180 | 125 | 55 | 0 |
| 2 | section c | 0.60 (0.40) | 120 | 50 | 70 | 0 |

Iteration 2: Required: 1000, Allocated: 875, Remaining Required: 125, Available in Sections: 0

**Algorithm 1** K-Maximal Word Allocation

**Inputs:**

$S_w \leftarrow list\_of\_section\_weights$
$W \leftarrow list\_of\_number\_of\_words\_in\_each\_section$
$K \leftarrow required\_number\_of\_words\_in\_final\_summary$
$K_{Alloc} \leftarrow list\_of\_allocated\_number\_of\_words\_to\_each\_section\_till\_previous\_iterations$

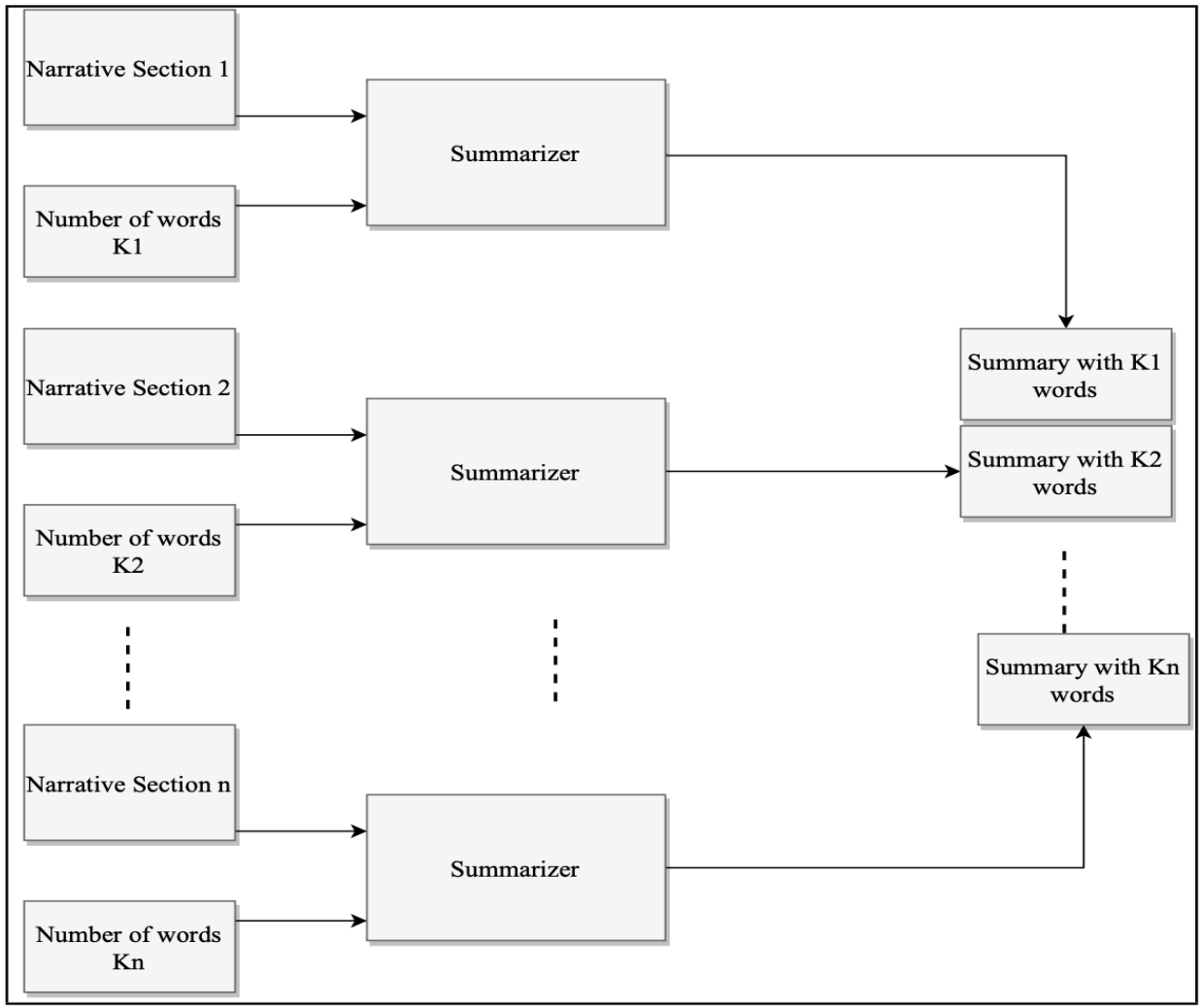
**procedure** ALLOCATE_MAXIMAL_WORDS

**if** $K = 0$ or $sum\_of(S_w) = 0$ **then**
**return** $K_{Alloc}$
**end if**
$S_w\_normalized = \mathrm{S}_w / sum\_of(S_w)$
$W_{Req} = \mathrm{K} \times S_w\_normalized$
**if** $W_{Req} \leq W$ **then**
**return** $K_{Alloc} + W_{Req}$
**else**
**return** $K_{Alloc} + W$
**end if**
**for** $i = 0$ to $length\_of(S_w)$ **do**
**if** $W_{Req}[i] >= W[i]$ **then**
$K_{Alloc}[i] = K_{Alloc}[i] + W[i]$
$K = K - W[i]$
$W[i] = 0$
**else**
$K_{Alloc}[i] = K_{Alloc}[i] + W_{Req}[i]$
$K = K - W_{Req}[i]$
$W[i] = W[i] - W_{Req}[i]$
**end if**
**end for**
$allocate\_maximal\_words(S_w, W, K, K_{Alloc})$
**end procedure**

# Distributed Summary Generation

- Information to be summarized is spread in different aspects which have their own importance and contribution towards summary.

- Once we identify the narrative sections and their contributions in terms of number of words, any summarizer can be used to generate summary



# Results

- We experimented with few summarizers
  - Top-K
  - BERT Extractive Summarizer
  - BART Extractive Summarizer

| Summarizer | R1P | R1R | R1F | R2P | R2R | R2F |
|---|---|---|---|---|---|---|
| BART | 0.544 | 0.444 | 0.417 | 0.304 | 0.244 | 0.232 |
| BERT | 0.56 | 0.40 | 0.42 | 0.32 | 0.20 | 0.22 |
| Top-k | 0.523 | 0.596 | **0.508** | 0.347 | 0.418 | **0.345** |

Table 6: Comparision of Summarizers for Generating Distributed Summary on English Validation Dataset

# Official Results

- Ranked According to ROUGE-2 F1 Score.
- Overall Score: English (50%), Spanish (25%) and Greek (25%)

| Team | English | Spanish | Greek | Overall Score |
|---|---|---|---|---|
| LSIR-1 | 0.365 | **0.157** | 0.141 | 0.257 |
| **SSC-AI-RG-1** | 0.327 | **0.146** | **0.185** | 0.24625 |
| **SSC-AI-RG-3** | 0.319 | **0.146** | **0.185** | 0.24225 |
| IIC | 0.366 | 0.125 | 0.095 | 0.238 |
| **SSC-AI-RG-2** | 0.3 | **0.146** | **0.185** | 0.23275 |
| Team-Tredence-2 | 0.322 | 0.131 | 0.138 | 0.22825 |
| Team-Tredence-1 | 0.317 | 0.131 | 0.138 | 0.22575 |
| LIPI | 0.374 | 0.07 | 0.046 | 0.216 |
| Team-Tredence-3 | 0.322 | 0.131 | 0.072 | 0.21175 |
| LSIR-3 | 0.275 | 0.138 | 0.13 | 0.2045 |
| MACQUARIE-1 | 0.303 | 0 | 0 | 0.1515 |
| MACQUARIE-3 | 0.302 | 0 | 0 | 0.151 |
| MACQUARIE-2 | 0.301 | 0 | 0 | 0.1505 |
| AO-LANCS | 0.143 | 0.134 | 0.131 | 0.13775 |

# Future Work

- In this work, we focused on the inputs and outputs of summarizers. In future work, we would like to explore more sophisticated approaches for summarization using the foundations of using K- Maximally Allocated Words and Distributed Summary Generation.

- Our current approach is also dependent upon the TOC in English Reports. Alternate approaches need to be explored to reduce this dependency.