A Compare-and-contrast Multistage Pipeline for Uncovering Financial Signals in Financial Reports

Jia-Huei Ju[1], Yu-Shiang Huang[1,2], Cheng-Wei Lin[1], Che Lin[2,3,4], and Chuan-Ju Wang[1,2]

[1]Research Center for Information Technology Innovation, Academia Sinica,
[2]Graduate Program of Data Science, National Taiwan University and Academia Sinica,
[3]Graduate Institute of Communication Engineering, National Taiwan University,
[4]Department of Electrical Engineering, National Taiwan University

Table of Contents

Introduction

For financial practitioners, financial report is one of the most important materials for knowing a company's operation. For example, the Form 10-K is

- mandated: required by the SEC.
- periodically released
- publicly available
- **comprehensive**: contains full description of a company's financial activities.

NVIDIA CORPORATION
TABLE OF CONTENTS

These documents are so informative; however, mining useful signals needs lots of human efforts.

Introduction: Motivations

We observe that financial corpus is

1. High overlapping characteristics: on average, about **80% of tokens** used in a company's reports are the **same** (except the "date").
2. Yearly-dependant: contents are much **more similar** between arbitrary **adjacent years** than the distant one.

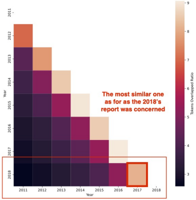


Figure 1: Text similarity heatmap of used tokens between years (from 2011 to 2018). The blocks with lighter color indicate there are more similar.

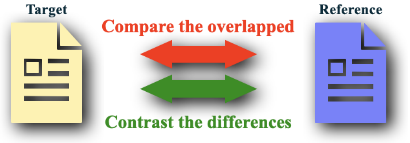
Based on these characteristics, we introduce a **highlighting task** and proposed a **multistage pipeline** to address the empirical problems.

Problem/Task Definitions

Definitions: The Highlighting Task

The reference-to-target structure:

- **Target** ($\mathcal{D}_\ell$): a focal financial report at year ℓ.
- **Reference** ($\mathcal{D}_{\ell-1}$): the same company's report at year $\ell - 1$.
- A document pair contains **multiple reference-to-target** (t, r) **segment pairs**; we denote them as a set $\mathcal{T}$.[1]



Highlighter f have to predict the underlying **rationale/important words** by comparing and contrasting the contexts of a given sentence pair.

[1] Note that we filter some *irrelevant* (t, r) pairs using a heuristic manner to relieve the human evaluation burden.

Definitions: The Highlighting Task (example)

The highlighting task

$$\mathbf{R} \triangleq P_f(t|r), \quad t \in \mathcal{D}_\ell, r \in \mathcal{D}_{\ell-1}$$

- **R**: the **rationale (words)** of the relations of a given (t, r) pair.
- $P_f(\cdot)$: the **word importance** predicted by a highlighting model f.

The words with higher importance are regarded as **financial signals**.[1]

$\mathcal{T}^\alpha$	2017 (reference)	*Net sales in the Americas* **increased 5%**, *or $201.8 million, to $4,302.9 million...*
	2018 (target)	*Net sales in the Americas* **decreased 1%**, *or $58.5 million, to $4,513.8 million...*

Table 1: An example of reference-to-target pair.

[1] There are still many factors affect what should be considered as signals; we have a brief discussion in Limitation in our paper.

The Multistage Pipeline

Proposed Pipeline: Overview

Our pipeline design includes the following stages:

- S_0 – Document segmentation
- S_1 – **Relation Recognition**
- S_2 – **Out-of-domain Fine-tuning &** S_{2_+} **– In-domain Fine-tuning**

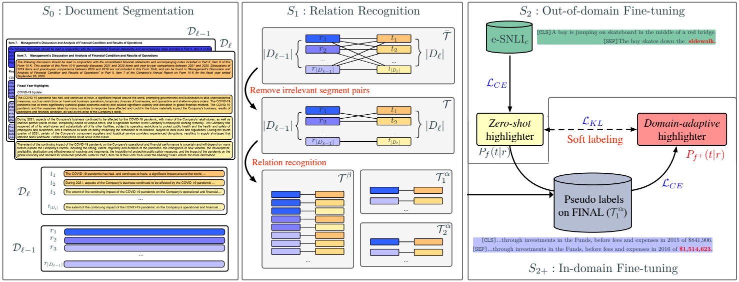


Figure 2: The compare-and-contrast multistage pipeline

Proposed Pipeline S_1: Relation Recognition

After document segmentation, we categorized each reference-to-target segment pairs $(r, t) \in \mathcal{T}$ into:

- Insignificant relations ($\mathcal{T}^{\beta}$): uninformative, e.g. regulations.
- **Revised relations** ($\mathcal{T}_1^{\alpha}$): differ in few words but disclose different meanings, e.g., increase $\implies$ decrease.
- **Mismatched relations** ($\mathcal{T}_2^{\alpha}$): mutually exclusive meaning, e.g., new policies.

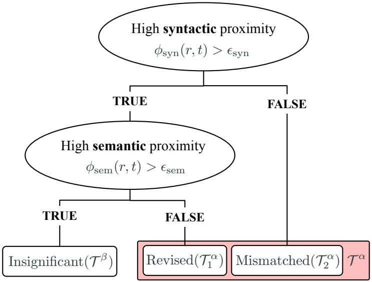
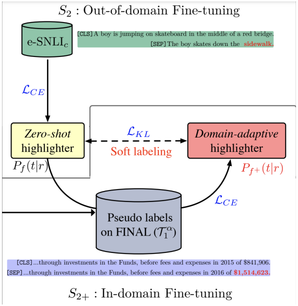


Figure 3: The heuristic filtering (categorization) procedure.

Proposed Pipeline S_2/S_{2_+}: Highlighting Stages

A two-staged fine-tuning approach for the **domain-adaptive** highlighter:

- Out-of-domain fine-tuning on e-SNLI$_c$ Train pairs.
- In-domain fine-tuning on the **Revised** pairs ($\mathcal{T}_1^{\alpha}$) with pseudo-labels.



Data		Example
e-SNLI$_c$	r	Children smiling and waving at camera
	t	The kids are frowning
Revised Pairs	r	Net sales in the Americas increased 5%, or $201.8 million ...
	t	Net sales in the Americas decreased 1%, or $58.5 million ...

Table 2: Example of the training pairs in S_2 and S_{2_+}. The words in red means the negative; the highlighted words are positive, and the other words are None.

Proposed Pipeline S_2/S_{2_+}: Highlighting Stages

As we transform the highlighting task into a **binary token classification task**, we can have models learn from the following objective functions:

Two-staged Fine-tuning

(S_2 Out-of-domain) Zero-shot highlighter f: (w/ e-SNLI$_c$)

$$\mathcal{L}_{\mathrm{CE}} = \sum_j -\left(Y_t^j \log P_f^j(t|r)\right) + \left(1 - Y_t^j\right) \log\left(1 - P_f^j(t|r)\right)$$

(S_{2_+} In-domain) Domain-adaptive highlighter f^+: (w/ pseudo-labels)

$$\mathcal{L}_{\mathrm{KL}} = \sum_j -\mathrm{KL}\Big(\underbrace{P_f^j(t|r)}_{Prior} \| P_{f^+}^j(t|r)\Big)$$

$$\mathcal{L}_{\mathrm{SL}} = \gamma \mathcal{L}_{\mathrm{CE}} + (1-\gamma)\mathcal{L}_{\mathrm{KL}}$$

Empirical Data and Evaluation

Evaluation: Datasets and Metrics

Evaluation dataset for highlighting task

e-SNLI$_c$ (Contradiction pairs)					
	#Pairs	Avg. $\|t\|$	Avg. $\|r\|$	Avg. $\#w_+$	Avg. $\#w_-$
Train	183,160	8.2	14.1	2.0	6.2
Test	3,237	8.1	15.3	2.1	6.0
FINAL (**FIN**ancial **AL**pha) Dataset					
	#Pairs	Avg. $\|t\|$	Avg. $\|r\|$	Avg. $\#w_+$	Avg. $\#w_-$
Train ($\mathcal{T}_1^\alpha$)	30,000	31.3	33.2	3.7	60.8
Eval ($\mathcal{T}_1^\alpha$)	200	33.2	31.3	5.5	25.9
Eval ($\mathcal{T}_2^\alpha$)	200	29.6	29.0	11.0	18.0

Table 3: Statistics of e-SNLI$_c$ and FINAL datasets.

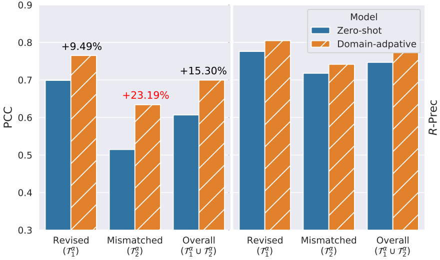
Evaluation metrics (R-prec: discrete; PCC: continuous)

- R-Prec: $\#$(top-R important words $\cap$ Annotated words)$/R$
- PCC: Pearson Correlation Coefficient (Predictions, Avg.annotation)

Evaluation: Highlighting Performance

Domain-adaptive highlighting models (# 4) outperform all the other settings and without lossing the generality of token representations.

#	W.U.	Labeling		FINAL		e-SNLI$_c$	
		P	**S**	*R*-Prec	PCC	*R*-Prec	PCC
Zero-Shot							
1	✓	✗	✗	0.7469	0.6067	0.8565	0.7555
Pseudo few-shot							
2	✗	✓	✗	0.6968	0.6368	0.6302	0.5752
Domain-adaptive							
3	✓	✓	✗	0.7160	0.6555	0.8475	0.7305
4	✓	✓	✓	**0.7865***	**0.7290***	**0.8605**	**0.7566**



Conclusion & Future Works

Conclusion and Future Works

This work

- A Financial signal highlighting **task**.
- A human-annotated **evaluation dataset**.
- A **multistage pipeline** with the domain-adaptive learning (S_2/S_{2_+})

Many possible future works include

- More effective: **financial corpus is abundant**; it is possible to pre-train a financial language models.
- More features: the **bi-directional** rationalization task; applying on other languages than English.
- More efficient: practitioners would like to explore more **end-to-end** way as an application, e.g., dense retrieval, explanation, etc.
- More modality: analyzing charts, tables, or cross-company, cross-sectors, etc.

Thank You!

Are there any questions you'd like to ask?

Jia-Huei Ju	jhjoo@citi.sinica.edu.tw
Yu-Shiang Huang	yushuang@citi.sinica.edu.tw
Cheng-Wei Lin	lcw.1997@citi.sinica.edu.tw
Che Lin	chelin@ntu.edu.tw
Chuan-Ju Wang	cjwang@citi.sinica.edu.tw