Multi-stage Pipeline for Discovering Rationales of Financial Signal

Anonymous ACL submission

1 Introduction

With the rapid growth in information, many tasks are focused on boosting the efficiency of comprehending information, especially in the natural language processing field. For example, summarization focuses on selecting or generating a subset of sentences that best represents the document's summary. Passage retrieval tries to select the most relevant passages given the query that the users care about. Reading comprehension gives attention to answering questions given massive documents. These tasks aim at mining important signals and saving human time when comprehending important contents from large texts.

In the finance application, understanding regulatory financial reports especially meet the scenario of efficiently mining the important signal in a huge amount of text. They contain rich information about the specific financial entity, and lots of works have shown mining information from financial reports could provide insightful signals to the users. For instance, Badertscher et al. (2018), Ertugrul et al. (2017) and You and Zhang (2009) indicate textual features from financial reports contain useful information about future firm performance and market reactions.

While these signals exist in the newly-published reports, the authority like SEC regulates that companies shall provide a comprehensive and detailed introduction of current status, making the reports often contain lots of unimportant and already-known information to the financial users. As figure 1 shows, the 10K forms in adjacent years have a high overlapping token ratio to each other, making it a lengthy process to absorb important signals in the new reports quickly. An intuitive solution is to treat the problem as a common-seen extractive summarization task in NLP field. However, there are no labeled reports-summary pairs data in the financial domain to meet the need.

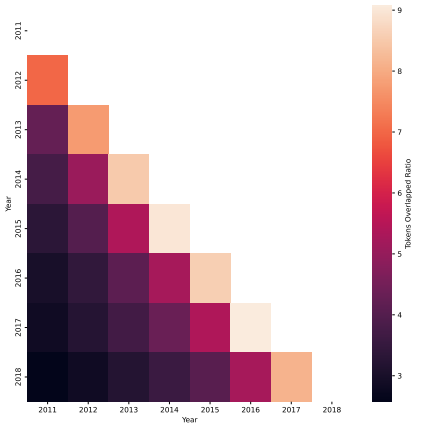

Figure 1: The overlapping token ratio between 10K forms item7 from 2011 to 2018

To automatically mine signals under the situation without labeled data, there are many attempts to discover rationales that are sensitive to variations in market measures. For instance, Kogan et al. (2009), Lin et al. (2021), and Tsai and Wang (2017) demonstrate the potential of leverage machine learning models to discover signals through the correlation of tokens and quantitative indicators from the financial market. Huang and Li (2011) and Lin et al. (2011) adapt clustering methods to distill information from the financial reports and automatically classify risk factors or stock price movements.

While these works rely on particular measures to discover rationales about important signals, the learned rationales depend on ad-hoc measures like risk volatility or stock returns and are counterintuitive. More specifically, when a new periodic financial report is published, the financial user would like to grasp important variations compared to what they already know quickly. The information they care about the most is the debut information about

the company and the material changes of existing entities instead of the signals related to specific tasks.

Upon this observation, we design a new task and a multi-stage pipeline to leverage the intrinsic year-to-year reference-to-target structure of such financial reports and highlight important signals. To be more specific, we leverage the last time published report as an information anchor and perform the following pipeline: First, we decompose the relation between 2 documents into segment relation pairs and reserve the relevant pairs. Second, we classify the pairs into 2 types to distinguish whether new information contains in the newly-published segments. Finally, we leverage pre-trained transformers and finetune them on a heuristic-labeled dataset from a class of pairs to obtain a model for highlighting signals. We conduct experiments on 200 companies' MD&A sections in SEC 10-K forms from 2011 to 2018 and provide a human-annotated evaluation dataset and select metrics close to the real-world application scenario to validate our proposed task and pipeline.

To sum up, our contributions are:

- We defined a new task that focuses on leveraging the time-to-time structure of periodic financial reports to help boost efficiency when comprehending new information for financial users.
- We demonstrate an efficient pipeline to find out the relevant segment relation pair and show the potential of leveraging domain transfer finetuning to highlight the signals hidden in the new-published periodic financial reports.
- We provide a human-annotated evaluation dataset (the FINAL dataset) for further research corresponding to highlight rationales in financial reports.

The paper is organized as follows: in section 2, we provide a comprehensive introduction of our proposed task. In section 3 we demonstrate our proposed pipeline. We introduce our evaluation dataset in section 4 and the corresponding experiment results are shown in section 5, while the ablation studies is provided in section 6.

2 Problem Definition

The inherent *year-to-year* property of financial reports enable us to leverage the *difference* between documents of a company in consecutive years. As the underlying *differences* sometimes disclose complex relationship within a pair of documents, in this paper, we aim to deep dive into such differences and further discover the relationship through rationales, which are finally viewed as important signals in financial reports. To this end, we first formally describe the concept and the related notations for the *reference-to-target structure* in Section 2.1; we then define the highlighting task in Section 2.2.

2.1 Reference-to-target Structure

Formally, for each company, we denote $\mathcal{D}_\ell$ as a set containing all segments in its financial report at year ℓ, where each element $d \in \mathcal{D}_\ell$ refers to a single segment. While we regard a focal company's financial report at year ℓ, $\mathcal{D}_\ell$, as the *target* document, we view the same company's report at year $\ell - 1$, $\mathcal{D}_{\ell-1}$, as the *reference* document. With such a year-to-year structure (i.e., the reference-to-target) of financial reports, we further break down the document-to-document relationship between $\mathcal{D}_\ell$ and $\mathcal{D}_{\ell-1}$ into enumerated segment-to-segment relationships, and we denote the set of enumerated segment pairs as $\bar{\mathcal{T}}$ hereafter.[1]

However, as $\bar{\mathcal{T}}$ includes all pairs of segments enumerated from $\mathcal{D}_\ell$, and $\mathcal{D}_{\ell-1}$ (i.e., $|\mathcal{D}_\ell||\mathcal{D}_{\ell-1}|$ pairs), it is intuitive that most of the segment pairs in it do not have worthwhile relationship to discover; simply put, for most segment pairs, the two segments are basically irrelevant. As a result, in this paper, we reduce the set $\bar{\mathcal{T}}$ to $\mathcal{T}$ by removing such off-the-subject segment pairs based on their syntactical similarities. Specifically, for each target segment $t \in D_\ell$, we first calculate the ROUGE-2 scores between the target segment t and all reference segments $r \in \mathcal{D}_{\ell-1}$ and sort the reference segments according to their scores in a descending order as

$$\bar{S}(t) = \left(r_1, r_2, \ldots, r_{|\mathcal{D}_\ell|}\right),^{2}$$

where r_k denotes the reference segment with the k-th highest ROUGE-2 score regarding the target segment t. With $\bar{S}(t)$, we then discard reference segments that fall behind the largest ROUGE-2 difference out of all possible ROUGE-2 differences,

[1] Note that each $\mathcal{D}_\ell$–$\mathcal{D}_{\ell-1}$ pair corresponds to a set of segment pairs $\bar{\mathcal{T}}$; for notation simplicity, we however do not use the subscript for $\bar{\mathcal{T}}$ to characterize the different sets.

[2] Note that we use round parentheses to represent the ordered set.

2015 (Ref.)	*Our most critical accounting policies relate to revenue recognition, inventory, pension and other post-retirement benefit costs, goodwill, other intangible assets and long-lived assets and income taxes.*
2016 (Target)	*Our most critical accounting policies relate to revenue recognition, inventory, pension and other post-retirement benefit costs, goodwill, other intangible assets and long-lived assets and income taxes.*

(a) An example of segment pairs in $\mathcal{T}^{\beta}$

2017 (Ref.)	*The Partnership experienced a net trading gain, through investments in the Funds, before fees and expenses in 2015 of $841,906.*
2018 (Target)	*The Partnership experienced a net trading loss, through investments in the Funds, before fees and expenses in 2016 of* ***$1,514,623.***

(b) An example of segment pairs in $\mathcal{T}^{\alpha}$

resulting in a truncated set S_t. Note that the remaining reference segments in truncated set is often below 5. Finally, with S_t, the reduced segment pair set is

$$\mathcal{T} = \{(t, r) | (t, r) \in \bar{\mathcal{T}} \wedge r \in S(t)\} \quad (1)$$

As in this study, we aim at locating meaningful financial signals revealed by the differences between each segment pair, we further classify each segment pair (t, r) in $\mathcal{T}$ into the following two sets:

1. $\mathcal{T}^{\beta} \subset \mathcal{T}$ refers the set containing target-reference segment pairs that possess vastly similar meaning (e.g., the mandated declaration in the same company's 10-K fillings, see Table 1a for an example). Generally, it can be said that there is no additionally noteworthy content in target segment t compared to reference segment r.

2. $\mathcal{T}^{\alpha} = \mathcal{T} \setminus \mathcal{T}^{\beta}$ refers the set containing segment pairs with dissimilar meanings (e.g. the changes of net sales, see Table 1b for an example). For the pairs in $\mathcal{T}^{\alpha}$, we further classify them into two types based on their syntactic and semantic proximity, the details of which are provided in Section 3.2.

2.2 Highlighting Task

In this work, we consider pairs in $\mathcal{T}^{\alpha}$ as the pairs of interest and seek to provide rationales by predicting the word importance for each segment pair $(t, r) \in \mathcal{T}^{\alpha}$ as

$$\mathbf{R} \triangleq P_f(t|r), \quad (2)$$

where $\mathbf{R}$ indicates the word importance of a target segment t conditioned on reference segment r. The highlighting model is denoted as f, which is detailed in Sections 3.3 and 3.4.

3 Proposed Pipeline

In this section, we detail the proposed multi-stage pipeline for discovering the rationale behind the reference-to-target structure in financial reports (see Figure 2). Our pipeline is composed of three parts: 1) the document segmentation stage S_0 in Section 3.1, 2) the relation recognition stage S_1 in Section 3.2, and 3) the stages of highlighting (S_2 and S_2^+) in Sections 3.3 and 3.4.

3.1 Document Segmentation

Financial reports are multi-modal, where they often cover multiple aspects and topics; generally, each aspect or topic may involve more than one consecutive sentences (usually one to three sentences) to convey its meaning. Therefore, instead of considering *sentences* as the basic unit of text, in this study, we regard the *uni-modal segments* as the smallest unit for constituting financial documents, for which we utilize the fine-tuned Cross-Segment BERT (Lukasik et al., 2020) to obtain coherent *uni-modal segments*. Note that some studies also shown that breaking a document into uni-modal segments is beneficial to downstream applications (Shtekh et al., 2018; Qiu et al., 2022).

3.2 S_1: Relation Recognition

With the inherent *reference-to-target* structure of financial reports introduced in Section 2.1, in this stage, we further propose a systematic procedure to manage the two types of relations $\mathcal{T}^{\beta}$ and $\mathcal{T}^{\alpha}$ with semantic and syntactic proximity. Specifically, we use two proximity functions in our procedure, the ROUGE-2 and the Sentence-BERT (Reimers and Gurevych, 2019) cosine similarity[3], for assessing the syntactic and semantic proximity between each reference-to-target pair $(t, r) \in \mathcal{T}$. The scores for the syntactic and semantic proximity are dented as $\phi_{\text{syn}}(t, r)$ and $\phi_{\text{sem}}(t, r)$, respectively.[4]

We disentangle the complex relations by jointly considering the above two scores of each segment pair, the procedure of which is summarized in Figure 3. Specifically, we empirically design a rule-

[3]We adopt the `MEAN` pooling strategy to derive segment embeddings.

[4]Note that the scoring functions are not limited to the two but can be replaced to others suitable for the task.

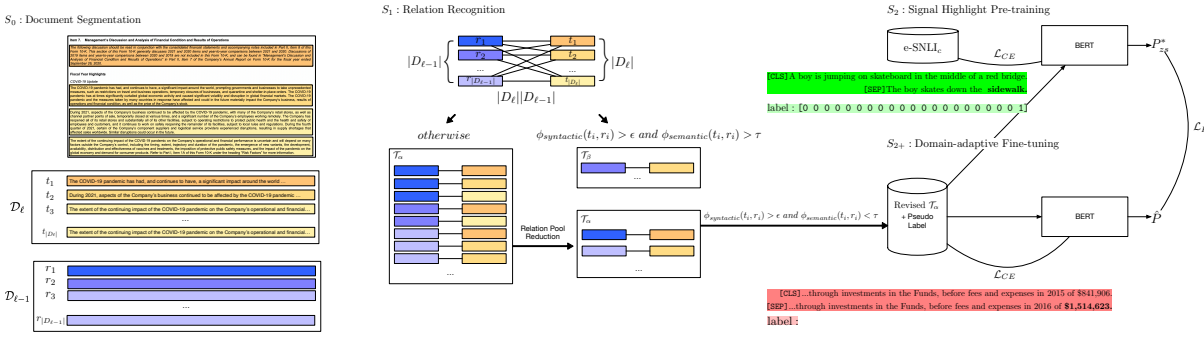

Figure 2: The proposed pipeline for discovering rationales.

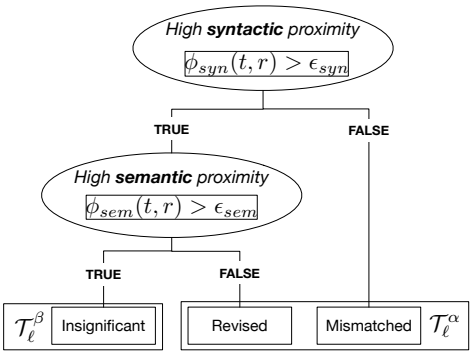

Figure 3: Procedure diagram of relation recognition

based procedure and classify each segment pair into three types as follow.

1. *Insignificant* relations ($\mathcal{T}^\beta$) reveal uninformative segment pairs with highly similar syntactic and semantic meanings between the target and reference segments (i.e., $\phi_{\text{syn}} > \epsilon_{\text{syn}}$ and $\phi_{\text{sem}} > \epsilon_{\text{sem}}$).

2. *Revised* relations ($\mathcal{T}_1^\alpha$) reveal segment pairs differ in some words only but disclose quite different meanings, resulting in a high $\phi_{\text{syn}}(t, r)$ but a relatively low $\phi_{\text{sem}}(t, r)$ (i.e., $\phi_{\text{syn}} > \epsilon_{\text{syn}}$ and $\phi_{\text{sem}} < \epsilon_{\text{sem}}$).

3. *Mismatched* relations ($\mathcal{T}_2^\alpha$) indicate the meanings of segment pairs are to some extent mutually exclusive, and thus resulting in a low $\phi_{\text{syn}}(t, r)$ (i.e., $\phi_{\text{syn}} < \epsilon_{\text{syn}}$).

Note that the setting of the two thresholds ϵ_{syn} and ϵ_{sem} is discussed in Section 5. In the following subsections, we mainly focus on discovering financial signals concealed in the last two types of relations, $\mathcal{T}^\alpha = \mathcal{T}_1^\alpha \cup \mathcal{T}_2^\alpha$ (see the definition in Section 2.2).

3.3 S_2: Out-domain Fine-tuning

In this stage, we aim to pinpoint financial signals for segment pairs in $\mathcal{T}^\alpha$. Specifically, as defined in Section 2, we focus on discovering alignment rationales with important words in the target segment t given a target-reference segment pair (t, r), where the rationales are inferred conditioned on the reference segment r (see Eq. (2)).

Binary token classification task To this end, we cast the word importance prediction into a binary token classification task, and thus, we develop highlighting models by means of supervised learning. Specifically, we first leverage the pre-trained BERT (Devlin et al., 2019) to construct contextualized reference-to-target pair representation, where each pair of interests (i.e., $(t, r) \in \mathcal{T}^\alpha$) constitutes an input with the special tokens as

$$\mathbf{h}_{(t,r)} = \text{BERT}(\texttt{[CLS]}\, r\, \texttt{[SEP]}\, t),$$

where $\mathbf{h}_{(t,r)} \in \mathbb{R}^{n \times d}$ indicates the contextualized tokens representation of the pair. Note that above d denotes the dimension of each token representation, and n is the number of tokens in (t, r). Secondly, on top of the tokens representation $\mathbf{h}_{(t,r)}$, we add a highlighting model f (a MLP layer) with a softmax activation function. As a result, the conditional word importance $P_f^{(j)}(t|r)$ for the j-th word in target segment t is

$$P_f^{(j)}(t|r) = \frac{f^{(1)}\left(\mathbf{h}_{(t,r)}[j]\right)/\tau}{\sum_{i=1}^{2} f^{(i)}\left(\mathbf{h}_{(t,r)}[j]\right)/\tau}, \quad (3)$$

where $\mathbf{h}_{(t,r)}[j]$ denote the token representation of the j-th node in target segment t (i.e., the j-th row vector of $\mathbf{h}_{(t,r)}$), $f(\cdot) : \mathbb{R}^d \to 2$, and τ is a hyper-parameter that controls the probability distribution.

Signal highlighting warm-up As we regard the signal highlighting task as a binary token classifi-

cation task, we first fine-tune the model f on an external human-annotated dataset, e-SNLI (Camburu et al., 2018), to obtain a *zero-shot* model. Note that the e-SNLI dataset is compiled for explanation generation task for distinguishing relations of aligned sentence pairs (t', r') (i.e,. premise and hypothesis) in natural language inference (NLI). We then treat the human-annotated rationale words as the ground-truth important words for the *premise-to-hypothesis* relation. It is worth mentioning that the pair structure is similar to our defined *reference-to-target* structure[5] in finance. Formally, we adopt the binary cross-entropy objective on each token in premise t' to fine-tune the BERT token representations and the highlighting model f as

$$\mathcal{L}_{\text{CE}} = \sum_{j} -\left(Y_{t'}^{(j)} \log P_f^{(j)}\left(t'|r'\right)\right). + \left(1 - Y_{t'}^{(j)}\right) \log\left(1 - P_f^{(j)}(t'|r')\right), \quad (4)$$

where $Y_{t'}$ is a vector that each element $Y_{t'}^{(j)}$ indicates the binary label of word importance for the j-th word in the premise sentence t'. For instance, $Y_{t'}^{(j)} = 1$ implies the j-th word in t' is annotated as important word conditioned on the given hypothesis sentence r'. We therefore construct the out-domain *zero-shot* highlighting model by fine-tuning on e-SNLI *contradiction* pairs, which is regarded as a baseline to proceed with the following financial domain adaptation.

3.4 S_{2+}: In-domain Fine-tuning

Generally, for an application particularly in niche domains like finance, adopting models with *zero-shot* setting may not be effective enough in practical. Also, several studies have shown that language models gain low-quality performance when suffering from *domain shift* (Li et al., 2022) [more citations?]. To address this potential issue, we equip the proposed pipeline with an extra *in-domain fine-tuning* stage to enabling our highlighting model to well-adapt to the focal financial domain. Specifically, we propose the domain-adaptive financial signal highlighting model f_+, which is comprised of the following two parts: (1) pseudo labeling with revised segment pairs, and (2) further fine-tuning with soft labels.

Pseudo-labeling with revised segment pairs We introduce a simple yet effective pseudo-labeling approach for financial signal highlighting tasks derived from the *revised* segment pairs (i.e., $\mathcal{T}_1^{\alpha}$) collected from stage S_1 (see Section 3.2). Recall that such segment pairs differ in some words only but disclose quite different meanings. With such a property, we establish a heuristic labeling approach for pseudo labels of financial signals. Intuitively, we treat all the *revised* words in target segments$t$ as important words and mark them as positive, while we randomly sample other words from t as the negative ones.[6]

Further fine-tuning with soft labels Furthermore, to overcome the deficiency of the assertive binary pseudo labels in the previous part, we adopt the soft labeling technique to make token representations more generalized. Initially, as we illustrated in Figure 2, we leverage the *zero-shot* highlighting model f learned in stage S_2 (see Eq. (3)) to estimate the word importance of segment pairs in $\mathcal{T}_1^{\alpha}$, the results of which are regarded as soft labels () compared to assertive pseudo binary labels. Afterwards, we construct the soft-labeling objective $\mathcal{L}_{\text{SL}}$ as

$$\mathcal{L}_{\text{SL}} = \gamma \mathcal{L}_{\text{CE}} + (1 - \gamma)\mathcal{L}_{\text{KL}}, \quad (5)$$

where

$$\mathcal{L}_{\text{KL}} = -\text{KL}\left(P_f^{(j)}(t|r) \middle\| P_{f^+}^{(j)}(t|r)\right). \quad (6)$$

Above in Eqs. (5) and (6), $\text{KL}(\cdot)$ denotes the Kullback–Leibler (KL) divergence, and $P_f(t|r)$ and $P_{f^+}(t|r)$ indicate the (fixed) posterior probability and estimated probability distributions predicted by f and f_+, respectively. Finally, we finetune the highlighting model f_+ with the pseudo labels annotated on segments in $\mathcal{T}_1^{\alpha}$ via optimizing $\mathcal{L}_{\text{SL}}$ in Eq. (5). Note that we not only utilize probabilities $P_f(t|r)$ as our training targets (i.e., soft labels) for $\mathcal{L}_{\text{KL}}$ but also leverage the warm-start token representations of f as the initial checkpoint for fine-tuning f_+. In addition, we found that the hyperparameters, τ and γ, affect the performance significantly. Hence, we conduct experiments with hyperparameter search and determine the best practice as $\tau = 1$ and $\gamma = 0.1$, the detailed of which is reported in Section 6.

[5] In this work, we specifically select the *contradiction* pairs in e-SNLI as such a relationship is similar to our ultimate goal.

[6] We set the number of negative labels three times larger than that of the positive ones. In our experiments, we discover that fine-tuning only on fewer negative labels makes models more stable.

4 The FINAL Dataset

For the financial signal highlighting task, we build a synthetic dataset, named FINAL (**FIN**ancial-**AL**pha). FINAL is consisted of 30,000 reference-to-target segment pairs $(t, r) \in \mathcal{T}^{\alpha}$, where the target segments t covered underlying financial signals as we described in Section 3. The construction procedure includes: 1) Financial 10K corpus pre-processing, 2) Data generation and 3) Human judgement rationales, which are outlined below.

Financial 10-K corpus pre-procesing We use the financial Form 10-K filings collected from the Software Repository of Accounting and Finance, where a Form 10-K is an annual report required by the U.S. SEC. Firstly, we discard 10K fillings for which companies in some years are missing during 2011 to 2018. Out of these remaining 3,849 companies, we random sample 200 companies as the smaller subset of financial reports. For each reports, we take out solely the MD&A section (ITEM 7) as a standalone document $\mathcal{D}$, where mainly discuss important decisions, operational activities or sales performance. Finally, we align each document $\mathcal{D}_{\ell}$ with its corresponding last-year document $\mathcal{D}_{\ell-1}$, which result in 1400 *reference-to-target* document pairs. (200 companies $\times$ 7 year-to-year pairs).

Data generation Following our proposed multi-stage pipeline, we first pass each document pairs through stage S_0 to obtain enumerated set of segment pairs, and the smaller subset without off-the-topic pairs (i.e., $\mathcal{T}$ defined in Section 2). Next, we follow S_1 and the procedure illustrated in Figure 3 to classify the segment pairs into two groups: $\mathcal{T}_1^{\alpha}$ and $\mathcal{T}_2^{\alpha}$. From each of these two group, we randomly sample 200 pairs as our final 400 evaluation pairs. As for the training data, we randomly sample 30K pairs from the rest of revised segments pairs (i.e., $\mathcal{T}_1^{\alpha}$), and adopt the pseudo-labeling approach introduced in Section 3.4. The detail statistics of training and evaluation set are reported in Table 2a.

Human judged rationales To evaluate the empirical effectiveness, we conduct the human annotation process on 400 evaluation segment pairs. For each segment pairs (t, r), we collect human-annotated important words (i.e., the financial signals) from three university student annotators. Specifically, the annotators were asked to distinguish which words in target segment t shall be regarded as rationales according to the context reference segment r provides. That is, the annotated words should dictate the *reference-to-target* relationship or disclose extra information of interest[7], while the other words in target segment t were labeled as negative.

	# Examples	Avg.$\lvert t\rvert$	Avg.$\lvert r\rvert$	Avg.$\lvert w_+\rvert$	Avg.$\lvert w_-\rvert$
Train	30000	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

(a) Highlighting dataset of FINAL.

	# Examples	Avg.$\lvert t\rvert$	Avg.$\lvert r\rvert$	Avg.$\lvert w_+\rvert$	Avg.$\lvert w_-\rvert$
Train	183160	8.2	14.1	2.0	6.2
Test	3237	8.1	15.3	2.1	6.0

(b) Augment highlighting dataset of e-SNLI$_c$

Table 2: Highlighting dataset

We also assess the inter-rater reliability of three annotations with Fleiss' κ (Fleiss et al., 1971). For simplicity, we treat all the words in target segments as an independent classification task, which are 12K words among the target segments t in 400 evaluation pairs. Each subset of evaluation pairs have sufficient κ values from 0.6 to 0.7 in two types of relation, and 0.66 for overall. The other statistics of FINAL evaluation set are reported in Table 2a, including average word length of target and reference segments (denoted as $|t|$ and $|r|$), average number positive and negative words with in a target segment (denoted as $|w_+|$ and $|w_-|$). The FINAL dataset is also available in `https://github.com`.

5 Experiments

Sections 5.1 and 5.2 first describe the datasets and metrics for evaluating the highlighting performance, respectively; we then introduce the compared models in Section 5.3. Finally, we report the experimental results in Section 5.4.

5.1 Evaluation Datasets

FINAL As constructed in Section 4, FINAL is a financial signal highlighting dataset with segment pairs of interest (i.e., $\mathcal{T}_1^{\alpha}$ and $\mathcal{T}_2^{\alpha}$), where the target segment in each pair is with manually annotated rationale words. We then evaluated the highlighting performance on such binary-label word-level annotations.

[7]The annotation guideline is in Appendix **??**.

e-SNLI$_c$ We additionally evaluated the performance on e-SNLI, which is a natural language inference dataset with word-level human-annotated rationale. Particularly, in this paper, we only used the premise-to-hypothesis sentence pairs labeled as *contradiction* (denoted as e-SNLI$_c$) in the test set of the e-SNLI datatset for evaluation.

5.2 Evaluation Metrics

Recall-sensitive fraction of signals In practice, financial practitioners usually concern more about the recall of founded signals rather than their precision due to the high cost of missing signals. Accordingly, we borrow the idea of the metric in the field of information retrieval—the R-precision (Buckley and Voorhees, 2000) for evaluation; in our case, R-precision (denoted as R-Prec hereafter) is the precision at R, where R is the number of annotated words in each target segment; simply put, if there are r annotated words among the top-R predicted words, then R-precision is r/R.

Sequence agreement of word importance In addition, we attempt to measure the agreement between the predicted importance of words for each target segment (considered as a consecutive number sequence) and its corresponding ground-truth sequence.[8] Specifically, we utilized the Pearson correlation coefficient (abbreviated as PCC hereafter) for evaluation.

It is worth mentioning that neither of the above two metrics need to set a hard threshold to determine the important words for evaluation. While R-Prec considers the words with top-R highest predicted probabilities, PCC directly leverages the predicted probabilities of words as the importance of words for calculation.

5.3 Compared Methods

Zero-shot We fine-tuned the BERT-base model on e-SNLI$_c$ training set with the binary token classification cross-entropy objective (See Section 3.3 for detail), and regarded such an approach as zero-shot with respect to financial signal highlighting.

Pseudo few-shot Instead of using e-SNLI$_c$, we fine-tuned the BERT-base model on the 30,000 revised segment pairs in $\mathcal{T}_1^{\alpha}$ (see the "Train" data in Table 2a) with the pseudo-labeled tokens (see pseudo labeling introduced in Section 3.4) and regarded such an approach as pseudo few-shot with respect to financial signal highlighting.

#	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	
Domain-adaptive							
3	✓	✓	✗	0.7160	0.6555	0.8475	0.7305
4	✓	✓	✓	**0.7865**	**0.7290**	**0.8605**	**0.7566**

Table 3: Highlighting performance

Domain-adaptive With the use of the *zero-shot* highlighting model, we further conducted the in-domain fine-tuning (see stage S_2^+ in Section 3.4) for domain adaptation.

For fair comparison, all the above models share the same training setups such as the batch size, learning rate, and optimizer.[9]

5.4 Empirical Results

5.4.1 Main Results for Signal Highlighting

Performance on FINAL Table 3 tabulates the highlighting performance under four conditions (i.e., #1–#4), where W.U. denotes that e-SNLI$_c$ is used for warm-up fine-tuning (i.e., the zero-shot highlighting model), **P** and **S** denote the pseudo and soft labeling, respectively.

We first focus on the results of the main task on FINAL. As shown in the table, the proposed domain-adaptive approach with the use of both pseudo and soft labeling techniques (i.e., condition #4) achieves the best R-Prec of 0.7865 and PCC of 0.7290. In addition, from the significant performance increase from condition #2 to #3, we notice that *warm-up fine-tuning* (W.U.) plays an essential role on such a financial signal highlighting task. Similarly, the *soft labeling* technique is also beneficial to our task, which brings about 10% performance improvement in terms of both evaluation metrics (by comparing the results of conditions #3 and #4). On the other hand, from the results of conditions #1 and #3, we observe that solely adopting the pseudo labeling might not be helpful for the signal highlighting task. This phenomenon may due to the fact that the pseudo labels constructed by

[8] Note that we here take the mean agreement of three human annotations as the ground truth to avoid suffering from subjective opinions of individuals.

[9] In this study, we adopted the batch size of 24 and fine-tuned/pre-trained the models for 2 epoches with BERT's default Adam optimizer and learning rate.

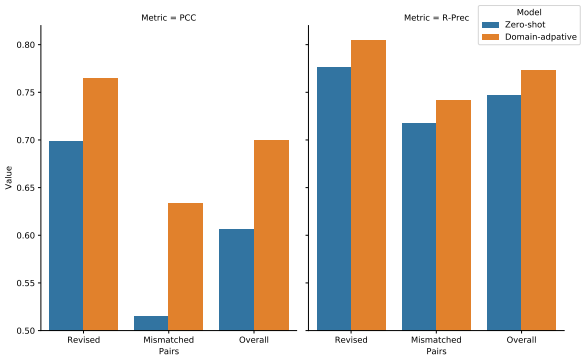

Figure 4: Highlighting effectiveness on different relations, including $\mathcal{T}_1^\alpha$ and $\mathcal{T}_2^\alpha$ with two our evaluation metrics.

	PCC		
	$\mathcal{T}_1^\alpha$ (revised)	$\mathcal{T}_2^\alpha$ (mismatched)	$\mathcal{T}^\alpha$
Zero-shot (#1)	0.6988	0.5146	0.6067
Domain-adaptive (#4)	0.7651	0.6339	0.6995
Improvement	9.49%	23.19%	15.30%

Table 4: Highlighting effectiveness on $\mathcal{T}_1^\alpha$ and $\mathcal{T}_2^\alpha$

the proposed heuristic approach (see Section 3.3) are too assertive on unimportant tokens, resulting in a biased highlighting model.

In sum, two main observations made from Table 3 are listed as follows.

- The proposed domain-adaptive fine-tuning approach with the pseudo and soft labeling techniques are effective for the signal highlighting task on financial reports.
- The warm-up fine-tuning and soft labeling are two crucial components to construct an effective domain-specific highlighting model.

Generalization ability between domains Table 3 also lists the results on the e-SNLI$_c$ testing data. Observed from the table, only the model with condition #4 performs on par with or even greater than the model with condition #1 (i.e., zero-shot), showing that the highlighting model fine-tuned by the our domain-adaptive approach with the pseudo and soft labeling techniques possess good generalization ability.

5.4.2 Analyses on Different Types of Target-reference Segment Pairs

To better understand the empirical advantages of our domain-adaptive approach, we further investigate the highlighting performance under different kinds of reference-to-target relations, $\mathcal{T}_1^\alpha$ (*revised*) and $\mathcal{T}_2^\alpha$ (*mismatched*). Figure 4 compares the results of the zero-shot (#1) and our domain-adaptive (#4) methods in terms of two metrics. However, we pay more attention to the PCC metric, which is less practical but more fine-grained as R-Precision only consider the set of important words (i.e., labeled as positive) instead of the whole words in target segment. In the left block of Figure 4, there is a significant difference between *revised* and *mismatched* relationship in terms of the PCC improvement. Particularly, the benefit of domain-adaptation on mismatched pairs is significantly greater than the one on the revised pairs, which brings approximately 23% PCC improvement. We hypothesized the important words in mismatched pairs are more uncertain which require intensive domain adaptation more than the words in revised pairs. It is worth noting that we only fine-tune the model on a small piece of revised segment pairs in $\mathcal{T}_1^\alpha$ for domain adaptation; however, the highlighting results of mismatched $\mathcal{T}_1^\alpha$ rise more significantly instead. The results suggests that our proposed domain-adaptive approach not only addresses the domain shift problem but also have a superior ability to infer the word importance even for the unfamiliar (unseen) relationship.

Reference Setup		FINAL		e-SNLI$_c$	
		R-Prec	PCC	R-Prec	PCC
Empty	[PAD]	0.4834	0.4033	0.6553	0.5687
Same	t	0.5108	0.3850	0.5697	0.4994
Random	r'	0.5345	0.4582	0.5658	0.4628
Original	r	0.7865	0.7290	0.8605	0.7566

Table 5: knowledge sources

Reference Setup		FINAL			
		Revised		Mismatched	
		R-Prec	PCC	R-Prec	PCC
Empty	[PAD]	0.3375	0.2780	0.6293	0.5286
Same	t	0.3872	0.3171	0.6343	0.4530
Random	r'	0.4164	0.3644	0.6527	0.5520
Aligned	r	0.8046	0.7651	0.7417	0.6339

Table 6: knowledge sources (option2)

6 Ablation Studies

To further understand what are the impacts of our pipeline setups with respect to final highlighting results, we conduct several ablation experiments and report our findings below.

Pseudo Labeling	FINAL		e-SNLI$_c$	
	R-Prec	PCC	R-Prec	PCC
Lexicon-based Labeling	0.6457	0.5774	0.6419	0.5847
+ Soft Label	0.6806	0.5932	0.8468	0.7261
Heuristic Labeling (#2)	0.6968	0.6368	0.6302	0.5752
+ Soft Label (#4)	0.7865	0.7290	0.8605	0.7566

Table 7: Lexicon labeling

Impact of Reference Knowledge Source In this ablation study, we aim to learn the impact of the *reference* segment, which is regarded as the knowledge context for discovering the financial signals in *target* segment. As we adopt the contextualized token representations of *reference-to-target* segment pair, we substitute the *original* reference segment (i.e., the most syntactically similar reference segment $r \in \mathcal{D}_{\ell-1}$ w.r.t target segment t) to other text and construct few variants of pairs for our highlighting model to inference. Specifically, we keep the target segment fixed but recast the BERT contextualized representation with the following variants with different reference segment:

- *Empty* indicates input with a single `[PAD]` token as reference segment, which implies *none* in BERT.
- *Same* means regarding the target segment as reference; in other word, this setting contains the duplicated target segments.
- *Random* indicates treating the arbitrary text as reference.

In the Table 5, we observe that *original* setups outperform the others in both FINAL and e-SNLI$_c$. However, the other setup perform relatively low in both of the evaluation dataset, which implies the context reference segments provided is essential. The high variance between setups also meet our assumption for financial signal highlighting, which we believe the important words may be diverse in some cases when the knowledge source (i.e., the reference segment) is differ.

Lexicon labeling with global importance With many studies have claimed the importance of financial lexicons (), we also leverage the external financial lexicons to strengthen our pseudo-labeling without supervised learning. Specifically, we use *Sentiment Word Lists* from Software Repository of Accounting and Finance to acquire the higher quality of positive and negative pseudo labels. We expand our pool of financial signal words (i.e., positive) with 3872 sentiment words from the wordlist.

Setting	τ	γ	FINAL	e-SNLI$_c$
Hard	1	1	0.7160	0.8475
Soft-mod	1	0.9	0.7200	0.8525
Soft-balance	1	0.5	0.7729	0.8574
Soft-tuned	1	0.1	0.7731	0.8606
Soft	1	0	0.7620	0.8626
Soft-smooth-0.5	0.5	0.1	0.7785	0.8613
Soft-smooth-2	2	0.1	0.7865	0.8605

Table 8: Domain-adaptive labeling

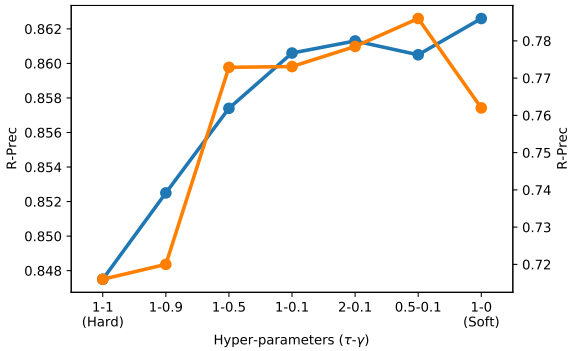

Figure 5: Domain-adaptive labeling (Option2)

By contrast, we use approximately 20K frequently-occurred words as the negative pool in combination with standard stopwords. Afterwards, we adopt the identical random sampling as our aforementioned labeling procedure in Section 3.4 as our *Lexicon-based Labeling* approach. In Table 7, we observe the deteriorated performance compared to heuristic labeling approach. Moreover, the *lexicon-based labeling* approach with soft labeling technique offset more than 10 points compared to our proposed domain adaptive approach (#4). We conclude that the sentiment words in lexicon represent the specific global importance among all financial reports. The global word importance is distant to our company-specific *reference-to-target* highlight tasks, which cares more about the local importance. Therefore, lexicon is not helpful and results in the poor effectiveness.

Impact of soft labeling To determine the better setup for training financial signal highlighting models with pseudo labels, we further explore the impact of hyperparameter τ and γ in terms of the practical metric R-Precision in FINAL and e-SNLI$_c$. As we have validated the effectiveness of our proposed *In-domain fine-tuning*, we adopt the same experimental setups for the later hyperparameter search. In Figure 5, we fine-tune few models with

different settings of hyperparameters, which the settings would compose distinct training objectives as we defined in Eq. (3) and Eq. (5). Specifically, we fine-tune models with solely *Hard* labels or *Soft* labels (i.e., the first and last setting in the figure.) as well as the other combinations of both in terms of cross-entropy objectives and KL divergence losses. We can observe that there exist a setting that can indeed maximize our R-Precision on FINAL, which is $\tau = 2, \gamma = 0.1$. However, in first and last setting in Figure 5, we observe that solely using the hard or soft labels can not achieve the highest effectiveness, which again emphasize the vital role of combining hard and soft labels for financial domain adaptation mentioned in Section 3.4.

7 Conclusion

TBD

References

Brad A Badertscher, Jeffrey J Burks, and Peter D Easton. 2018. The market reaction to bank regulatory reports. *Review of Accounting Studies*, 23(2):686–731.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 33–40, New York, NY, USA. Association for Computing Machinery.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mine Ertugrul, Jin Lei, Jiaping Qiu, and Chi Wan. 2017. Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis*, 52(2):811–836.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ke-Wei Huang and Zhuolun Li. 2011. A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Transactions on Management Information Systems (TMIS)*, 2(3):1–19.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of human language technologies: the 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 272–280.

Tian Li, Xiang Chen, Zhen Dong, Kurt Keutzer, and Shanghang Zhang. 2022. Domain-adaptive text classification with structured knowledge from unlabeled data. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4216–4222. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Ming-Chih Lin, Anthony JT Lee, Rung-Tai Kao, and Kuo-Tay Chen. 2011. Stock price movement prediction using representative prototypes of financial reports. *ACM Transactions on Management Information Systems (TMIS)*, 2(3):1–18.

Ting-Wei Lin, Ruei-Yao Sun, Hsuan-Ling Chang, Chuan-Ju Wang, and Ming-Feng Tsai. 2021. Xrr: Explainable risk ranking for financial reports. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 253–268. Springer.

Michal Lukasik, Boris Dadachev, Gonçalo Simoes, and Kishore Papineni. 2020. Text segmentation by cross segment attention. *arXiv preprint arXiv:2004.14535*.

Jielin Qiu, Jiacheng Zhu, Mengdi Xu, Franck Dernoncourt, Trung Bui, Zhaowen Wang, Bo Li, Ding Zhao, and Hailin Jin. 2022. Semantics-consistent cross-domain summarization via optimal transport alignment.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Gennady Shtekh, Polina Kazakova, Nikita Nikitinsky, and Nikolay Skachkov. 2018. Applying topic segmentation to document-level information retrieval. In *Proceedings of the 14th Central and Eastern European Software Engineering Conference Russia*, CEE-SECR '18, New York, NY, USA. Association for Computing Machinery.

Ming-Feng Tsai and Chuan-Ju Wang. 2017. On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1):243–250.

Haifeng You and Xiao-jun Zhang. 2009. Financial reporting complexity and investor underreaction to 10-k information. *Review of Accounting studies*, 14(4):559–586.