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

Anonymous ACL submission

Abstract

In this paper, we address the challenge of discovering financial signals in narrative financial reports. As these documents are often lengthy and tend to blend routine information with new information, it is challenging for professionals to discern critical financial signals. To this end, we leverage the inherent nature of the year-to-year structure of reports to define a novel signal-highlighting task; more importantly, we propose a compare-and-contrast multistage pipeline that recognizes different relationships between the reports and locates relevant rationales for these relationships. We also create and publicly release a human-annotated dataset for our task. Our experiments on the dataset validate the effectiveness of our pipeline, and we provide detailed analyses and ablation studies to support our findings.

1 Introduction

With the rapid growth of information, many tasks in the field of natural language processing (NLP) involve streamlining information comprehension. One such task is summarization, which selects a subset of sentences or generates new content that best represents the given document (Hermann et al., 2015; See et al., 2017; Cohan et al., 2018). This task helps humans save time and effort by identifying important information in a text. In the finance context, comprehending regulatory narrative reports is a classic example of efficiently mining signals from a large amount of text. As these reports often contain rich information concerning specific financial entities, discovering valuable insights is crucial for academia and the finance industry.

Much research has shown that textual features from financial reports contain valuable financial signals about future firm performance and market reactions (e.g., Badertscher et al., 2018; Ertugrul et al., 2017; You and Zhang, 2009). However, authorities such as the Securities and Exchange Commission (SEC) require that companies provide comprehensive and detailed information about their current status in these reports, which often contain much unimportant and already-known information. For example, the token overlap ratio between annual 10-K reports of the same company between adjacent years is often high,[1] making it a challenging and tedious task to acquire important signals in new reports (termed as the *overlapping characteristic* hereafter).

Recent advances in NLP technology have included attempts to efficiently and effectively comprehend lengthy financial documents. One approach to address this problem is through summarization (e.g., Zmandar et al., 2021b; Orzhenovskii, 2021; Gokhan et al., 2021). Other approaches additionally leverage numerical metrics, such as stock return volatility and abnormal trading volumes, to locate essential financial signals in reports (e.g., Kogan et al., 2009; Tsai and Wang, 2017; Rekabsaz et al., 2017; Agrawal et al., 2021; Lin et al., 2021). However, these approaches often require high-quality human annotation or suitable financial measures, which poses significant limitations in practical scenarios.

In this study, we approach financial report comprehension from a novel perspective by leveraging the intrinsic *year-to-year* characteristic of reports (i.e., the overlapping characteristics). Specifically, for a particular company, we use the document published in the previous year as an information anchor (i.e., the reference) to construct a year-to-year structure and locate important financial signals in the report of the subsequent year (i.e., the target). This inherent structure enables us to mine financial signals in a compare-and-contrast self-supervised manner, compared to existing supervised approaches.

Based on the year-to-year structure, we propose

[1] The overlap ratio calculated from Item 7 of the reports of the 3,849 companies from 2011 to 2018 (see FINAL dataset in Section 4) is around 0.826 on average.

a *compare-and-contrast multistage pipeline* to effectively locate financial signals in reports. We first identify a few types of relationships between reference and target financial reports at the segment level. Then, using these recognized relationships, we present a novel financial signal-highlighting task together with a domain-adaptive highlighting model. The goal of this task is to identify the rationales, represented by the importance of certain words, for a specific pair of year-to-year segments. Therefore, the words with high importance are deemed to be crucial financial signals in these reports. For experiments, we present a synthetic dataset consisting of 30,400 reference-to-target segment pairs for financial signal highlighting.[2] Experimental results validate the effectiveness of the proposed pipeline; detailed analyses and ablation studies are also provided.

2 Problem Definition

The year-to-year nature of financial reports allows us to take advantage of the differences between a company's documents in consecutive years. These differences may reveal complex but insightful relationships within a pair of documents. To better understand these relationships, we investigate them through rationales (represented by the word importance), which are considered essential signals in financial reports.

2.1 Reference-to-target Structure

Formally, for each company, $\mathcal{D}_\ell$ is 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. Given the annual nature (i.e., the reference-to-target structure) 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. We denote the set of enumerated segment pairs as $\bar{\mathcal{T}}$.[3]

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), intuitively, most segment pairs in $\bar{\mathcal{T}}$ have no interesting relationship. Hence, we reduce the set $\bar{\mathcal{T}}$ to $\mathcal{T}$ by removing irrelevant segment pairs based on their syntactical similarities. Specifically, for each target segment $t \in D_\ell$, we calculate the ROUGE-2 (Lin, 2004) 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 descending order as $\bar{S}(t) = \big(r_1, r_2, \ldots, r_{|\mathcal{D}_{\ell-1}|}\big)$.[4] With $\bar{S}(t)$, we then discard reference segments that fall behind the largest ROUGE-2 difference out of all possible ROUGE-2 differences, resulting in a truncated set $S(t)$.[5] Note that the difference is calculated between the two consecutive ROUGE-2 scores in $\bar{S}(t)$. Finally, with $S(t)$, the reduced segment pair set is $\mathcal{T} = \{(r,t) | (r,t) \in \bar{\mathcal{T}} \wedge r \in S(t)\}$.

(a) Segment pairs in $\mathcal{T}^\beta$

2017 (ref.)	*Our most critical accounting policies relate to revenue recognition, inventory, pension and other post-retirement benefit costs, goodwill, ...*
2018 (target)	*Our most critical accounting policies relate to revenue recognition, inventory, pension and other post-retirement benefit costs, goodwill, ...*

(b) Segment pairs in $\mathcal{T}^\alpha$

2017 (ref.)	*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: Segment pair classification

To locate meaningful financial signals revealed by segment pair differences, we further classify each pair $(r,t) \in \mathcal{T}$ into the following two sets:

1. $\mathcal{T}^\beta$ contains reference-to-target segment pairs with largely similar meanings (see Table 1(a)). Generally, there is no additionally noteworthy content in target segment t compared to reference segment r.
2. $\mathcal{T}^\alpha = \mathcal{T} \setminus \mathcal{T}^\beta$ contains segment pairs with dissimilar meanings (see Table 1(b)). Pairs in $\mathcal{T}^\alpha$ are further classified into two types based on their syntactic and semantic similarity, as discussed in Section 3.2.

2.2 Highlighting Task

We consider pairs in $\mathcal{T}^\alpha$ as the pairs of interest and provide rationales of underlying pairwise relationships by predicting the word importance for each

[2] The dataset and codes are available at https://anonymous.4open.science/r/fin_signal_highlighting/.

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

[4] The round parentheses represent the ordered set.

[5] Empirically, there are often one to five remaining reference segments in the truncated set for a target segment t.

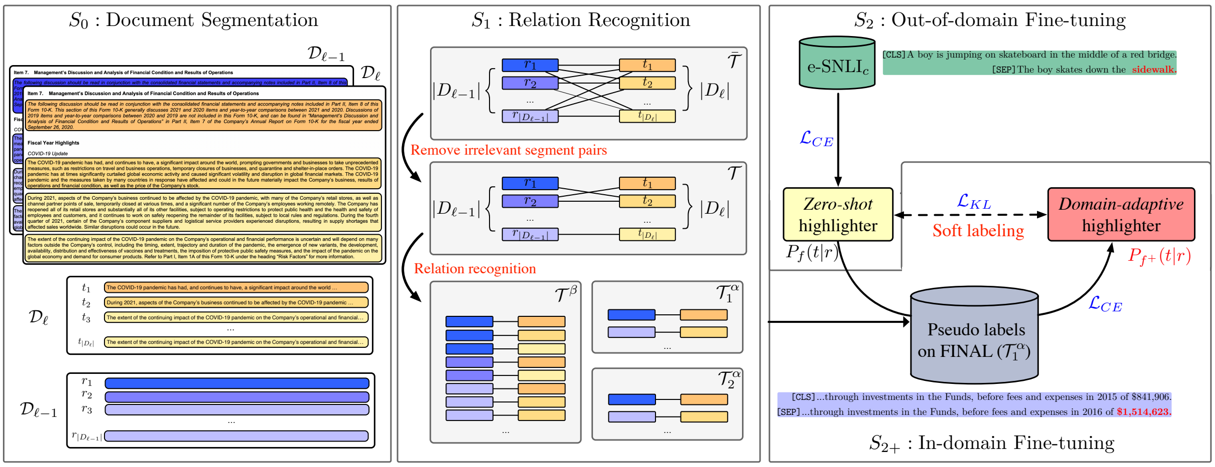
S_0 : Document Segmentation
$\mathcal{D}_{\ell-1}$
$\mathcal{D}_\ell$
Item 7. Management's Discussion and Analysis of Financial Condition and Results of Operations
Fiscal Year Highlights
COVID-19 Update
t_1 The COVID-19 pandemic has had, and continues to have, a significant impact around the world ...
t_2 During 2021, aspects of the Company's business continued to be affected by the COVID-19 pandemic ...
t_3 The extent of the continuing impact of the COVID-19 pandemic on the Company's operational and financial...
$t_{|D_\ell|}$
r_1
r_2
r_3
$r_{|D_{\ell-1}|}$
S_1 : Relation Recognition
$\bar{\mathcal{T}}$
$|D_{\ell-1}|$
$|D_\ell|$
Remove irrelevant segment pairs
$\mathcal{T}$
Relation recognition
$\mathcal{T}^\beta$
$\mathcal{T}_1^\alpha$
$\mathcal{T}_2^\alpha$
S_2 : Out-of-domain Fine-tuning
e-SNLI$_c$
[CLS] A boy is jumping on skateboard in the middle of a red bridge.
[SEP] The boy skates down the sidewalk.
$\mathcal{L}_{CE}$
Zero-shot highlighter
$\mathcal{L}_{KL}$
Soft labeling
Domain-adaptive highlighter
$P_f(t|r)$
$P_{f+}(t|r)$
$\mathcal{L}_{CE}$
Pseudo labels on FINAL ($\mathcal{T}_1^\alpha$)
[CLS] ...through investments in the Funds, before fees and expenses in 2015 of $841,906.
[SEP] ...through investments in the Funds, before fees and expenses in 2016 of $1,514,623.
S_{2+} : In-domain Fine-tuning

Figure 1: Proposed rationale discovery pipeline

segment pair $(r, t) \in \mathcal{T}^\alpha$ as

$$\mathbf{R} \triangleq P_f(t|r), \tag{1}$$

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

3 Proposed Pipeline

Here we describe the proposed multistage pipeline for discovering the rationale behind the reference-to-target structure in financial reports, as illustrated in Figure 1.

3.1 S_0: Document Segmentation

Financial reports are multimodal, often covering multiple aspects and topics; each aspect or topic usually uses one to three consecutive sentences to convey its meaning. Therefore, instead of considering sentences as the basic unit of text, we here regard *uni-modal segments* as the smallest unit for financial documents. We first use spaCy API for sentence segmentation.[6] Then, we utilize the fine-tuned cross-segment BERT (Lukasik et al., 2020) to obtain coherent uni-modal segments. Note that some studies show that breaking a document into uni-modal segments benefits downstream applications (Shtekh et al., 2018; Qiu et al., 2022; Chivers et al., 2022).

3.2 S_1: Relation Recognition

In this stage, a systematic procedure manages relation types $\mathcal{T}^\beta$ and $\mathcal{T}^\alpha$ with semantic and syntactic similarity. Specifically, we use two functions, ROUGE-2 and Sentence-BERT (Reimers and Gurevych, 2019) cosine similarity,[7] to assess the syntactic and semantic similarity between each reference-to-target pair $(r, t) \in \mathcal{T}$.[8] The scores for the syntactic and semantic similarity are denoted as $\phi_{\text{syn}}(r, t)$ and $\phi_{\text{sem}}(r, t)$, respectively.[9] We empirically design a rule-based procedure and classify each segment pair into three types.

1. *Insignificant* relations ($\mathcal{T}^\beta$) correspond to uninformative segment pairs with highly similar syntactic and semantic meanings between target and reference segment (i.e., $\phi_{\text{syn}} > \epsilon_{\text{syn}}$ and $\phi_{\text{sem}} > \epsilon_{\text{sem}}$).
2. *Revised* relations ($\mathcal{T}_1^\alpha$) correspond to segment pairs that differ in some words only but disclose quite different meanings, resulting in a high $\phi_{\text{syn}}(r, t)$ but a relatively low $\phi_{\text{sem}}(r, t)$ (i.e., $\phi_{\text{syn}} > \epsilon_{\text{syn}}$ and $\phi_{\text{sem}} < \epsilon_{\text{sem}}$).
3. *Mismatched* relations ($\mathcal{T}_2^\alpha$) correspond to segment pair meanings that are to some extent mutually exclusive, resulting in a low $\phi_{\text{syn}}(r, t)$ (i.e., $\phi_{\text{syn}} < \epsilon_{\text{syn}}$).

The procedure and the setting of the two thresholds (ϵ_{sem} and ϵ_{syn}) are also summarized in Figure 4 in Appendix C.

3.3 S_2: Out-of-domain Fine-tuning

Here we pinpoint financial signals for segment pairs in $\mathcal{T}^\alpha = \mathcal{T}_1^\alpha \cup \mathcal{T}_2^\alpha$. Specifically, for each segment pair $(r, t) \in \mathcal{T}^\alpha$, we discover rationales through predicted word importance in target segment t, where the rationales are inferred condi-

[6] https://spacy.io/api/sentencizer

[7] We derive segment embeddings using average pooling.

[8] Note that before the following procedure, we first reduce the set $\bar{\mathcal{T}}$ to $\mathcal{T}$ by removing irrelevant segment pairs (see Section 2.1).

[9] Note that the scoring functions are not limited to these two but can be replaced with other suitable functions.

tioned on reference segment r (see Eq. (1)).

Binary token classification To accomplish this, we cast the word importance prediction as supervised binary token classification. First, we leverage the pre-trained BERT (Devlin et al., 2019) model to construct contextualized reference-to-target pair representations, where each pair of interest constitutes an input with special tokens as

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

where $\mathbf{h}_{(r,t)} \in \mathbb{R}^{n\times d}$ is the contextualized token representation of the pair, d is the dimension of each token representation, and n is the number of tokens in segment pair (r, t). Second, on top of the token representation $\mathbf{h}_{(r,t)}$, we add a highlighting model $f(\cdot)$ (an MLP layer) with softmax activations. The resultant conditional word importance $P_f^j(t|r)$ for the j-th word in target segment t is

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

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

Signal highlighting warm-up As we view signal highlighting as binary token classification, we first fine-tune the model $f(\cdot)$ on e-SNLI (Camburu et al., 2018), an external human-annotated dataset, to obtain a zero-shot model. Note that e-SNLI was compiled for explanation generation with human-annotated rationales to distinguish relations of aligned sentence pairs (r', t') (i.e., premise and hypothesis) in natural language inference. We then treat the annotated words as the ground truth for the premise-to-hypothesis relation,[10] which is similar to our reference-to-target structure. Formally, we adopt the binary cross-entropy objective for each token in hypothesis t' to fine-tune the BERT token representations and the highlighting model $f(\cdot)$ as

$$\mathcal{L}_{\text{CE}} = \sum_j -\left(Y^j_{t'}\log P^j_f\left(t'|r'\right)\right.$$
$$\left. + \left(1 - Y^j_{t'}\right)\log\left(1 - P^j_f(t'|r')\right)\right),$$

where $Y_{t'}$ is a vector in which each element $Y^j_{t'}$ indicates the binary label of word importance for the j-th word in hypothesis t'. We thus construct the out-of-domain zero-shot highlighting model by fine-tuning on e-SNLI, which is regarded as a baseline to proceed with the following financial domain adaptation (see Figure 1).

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

Generally, for applications, particularly in niche domains like finance, models with a zero-shot setting may not be effective enough. Also, several studies show that language models exhibit poor performance under domain shift scenarios (Ben-David et al., 2006; Han and Eisenstein, 2019; Gururangan et al., 2020; Li et al., 2022). We account for this by equipping the proposed pipeline with an extra in-domain fine-tuning stage to enable our highlighting model to adapt properly to the financial domain. Specifically, we construct a domain-adaptive financial signal highlighting model $f_+(\cdot)$ with the following learning strategies: (1) pseudo-labeling with revised segment pairs in $\mathcal{T}_1^\alpha$, and (2) further fine-tuning with soft labels.

Pseudo-labeling with revised segment pairs We introduce a simple yet effective pseudo-labeling approach that uses revised segment pairs (i.e., $\mathcal{T}_1^\alpha$) collected from stage S_1 (see Section 3.2). Recall that these segment pairs differ in some words only but have quite different meanings. Given such a property, we establish a heuristic labeling approach for pseudo-labels of financial signals. Intuitively, we treat all revised words in target segment t as important words and mark them as positive, and randomly sample other words as negative ones.[11]

Further fine-tuning with soft labels To compensate for deficiencies in such assertive binary pseudo-labels, we use soft labeling to make the token representations more generalized. Initially, as illustrated in Figure 1, we leverage the zero-shot highlighting model $f(\cdot)$ learned at stage S_2 to calculate the approximate word importance of the revised segment pairs, the results of which are regarded as soft labels compared to the assertive pseudo-binary labels. We then construct the soft-labeling objective $\mathcal{L}_{\text{SL}}$ as

$$\mathcal{L}_{\text{SL}} = \gamma\mathcal{L}_{\text{CE}} + (1-\gamma)\mathcal{L}_{\text{KL}}, \tag{3}$$

[10] Here, we specifically select *contradiction* pairs in e-SNLI as this relationship is closer to our goal than the other two.

[11] We set the number of negative labels to three times that of the positive ones.

where

$$\mathcal{L}_{\text{KL}} = \sum_{j} -\text{KL}\left(P_f^j(t|r) \middle\| P_{f+}^j(t|r)\right) \qquad (4)$$

and γ is a hyperparameter that controls the impact of soft labeling. In Eqs. (3) and (4), $\text{KL}(\cdot)$ denotes Kullback–Leibler (KL) divergence, and $P_f(t|r)$ and $P_{f+}(t|r)$ indicate the estimated probability distributions predicted by $f(\cdot)$ and $f_+(\cdot)$, respectively. Finally, we fine-tune the highlighting model $f_+(\cdot)$ with the pseudo-labels annotated on segments in $\mathcal{T}_1^\alpha$ by optimizing $\mathcal{L}_{\text{SL}}$ in Eq. (3). 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 we also adopt the warm-start token representations and highlighting layer $f(\cdot)$ as the initial checkpoint for fine-tuning $f_+(\cdot)$. In addition, we discover that hyperparameters τ and γ affect the performance significantly. We discuss the hyperparameter search in Appendix B.

4 The FINAL Dataset

We constructed FINAL (**FIN**ancial-**AL**pha), a financial signal highlighting dataset, consisting of 30,400 reference-to-target segment pairs in $\in \mathcal{T}^\alpha$.

4.1 Financial 10-K Corpus Preprocessing

We used Form 10-K filings collected from the Software Repository for Accounting and Finance,[12] where a Form 10-K is an annual report required by the U.S. SEC. Specifically, we used 10-K filings from 2011 to 2018, which comprise 63,336 filings from 12,960 public companies. To make the best use of the year-to-year information, we discarded companies for which the reports in some years were missing during the period; 3,849 companies (3,849×8=30,792 reports total) remained after this filtering. We then randomly sampled 200 companies from the 3,849 companies with their annual reports to construct the dataset. In addition, while every 10-K annual report contains 15 schedules (e.g., Items 1, 1A, 1B, 2, 3, . . . , 7, 7A, . . . , 15), [13] we extracted only Item 7 (Management's Discussion and Analysis of Financial Conditions and Results of Operations ("MD&A")) to form the FINAL dataset.[14] Finally, we aligned each document $\mathcal{D}_\ell$ with its corresponding last-year document $\mathcal{D}_{\ell-1}$, resulting in 1,400 reference-to-target document pairs (i.e., 200 companies × 7 year-to-year pairs).

(a) FINAL dataset

	Pairs	Avg. $\lvert t\rvert$	Avg. $\lvert r\rvert$	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

(b) e-SNLI$_c$ dataset

	Pairs	Avg. $\lvert t\rvert$	Avg. $\lvert r\rvert$	Avg. $\#w_+$	Avg. $\#w_-$
Train	183,160	8.2	14.1	2.0	6.2
Test	3237	8.1	15.3	2.1	6.0

Table 2: Dataset statistics

4.2 Year-to-year Segment Pair Generation

After preprocessing, we followed the proposed multistage pipeline by first passing each document pair through stage S_0 to obtain an enumerated set of segment pairs $\bar{\mathcal{T}}$; we then reduced $\bar{\mathcal{T}}$ to $\mathcal{T}$ by removing irrelevant segment pairs (see Section 2.1). Next, we followed the relation recognition stage S_1 in Section 3.2 to obtain the two groups of segment pairs: $\mathcal{T}_1^\alpha$ and $\mathcal{T}_2^\alpha$. From each of these two groups, we randomly sampled 200 pairs for human annotation as our evaluation sets. Likewise, we randomly sampled 30,000 pairs from the rest of the revised segment pairs (i.e., $\mathcal{T}_1^\alpha$) as the training set for the pseudo-labeling approach in Section 3.4.

4.3 Human Annotation

To evaluate the empirical effectiveness of the proposed pipeline, we manually annotated the sampled 400 segment pairs. For each segment pair (r, t), we collected the labels of rationales from three annotators. Specifically, the annotators were to distinguish which words in each target segment t to regard as important financial signals according to the context of the corresponding reference segment r. That is, the words with positive labels were to characterize the reference-to-target relationship or disclose extra information of interest,[15] whereas the rest of the words in t were labeled as negative. We further assessed the inter-rater reliability of the three annotations with Fleiss' κ (Fleiss, 1971). For simplicity, we treat the prediction for the importance of each word in the target segment as independent classification tasks (containing roughly 12K words in the 400 evaluation pairs): for evaluation pairs from $\mathcal{T}_1^\alpha$, $\kappa = 0.71$; for those from

[12] https://sraf.nd.edu/sec-edgar-data/

[13] https://en.wikipedia.org/wiki/Form_10-K

[14] This setting follows most of the literature regarding textual analysis of financial reports.

[15] The annotation guidelines are provided in Appendix D.

$\mathcal{T}_2^\alpha$, $\kappa = 0.60$. The training and evaluation sets are described in Table 2(a), where Avg. $|t|$ and Avg. $|r|$ are the average lengths of target and reference segments, respectively, and Avg. $\#w_+$ and Avg. $\#w_-$ are the average numbers of words annotated as positive and negative, respectively.

5 Experiments

5.1 Evaluation Datasets

FINAL We evaluated the highlighting performance on the two evaluation sets with the human-annotated ground truth (see Table 2(a)).

e-SNLI$_c$ We additionally evaluated the performance on e-SNLI. Particularly, in this paper, we used only the premise-to-hypothesis sentence pairs labeled as *contradiction* (denoted as e-SNLI$_c$) in the test set of the e-SNLI dataset for evaluation (see Table 2(b)).

5.2 Evaluation Metrics

Recall-sensitive metric In practice, financial practitioners are usually concerned more about the recall of the discovered signals than their precision due to the high cost of missing signals. Accordingly, we borrow the idea of R-precision (Buckley and Voorhees, 2000), a metric from the information retrieval field. In our case, R-precision (R-Prec) is the precision at R, where R is the number of annotated words in each target segment: if there are r annotated words among the top-R predicted words, then the R-precision is r/R.

Sequence agreement of word importance In addition, we measure the agreement between the predicted importance of words for each target segment (considered as a number sequence) and its corresponding ground-truth sequence. Specifically, we use the Pearson correlation coefficient (PCC) for evaluation.

Note that for R-Prec, we use majority voting to derive single ground-truth labels from the three annotators, whereas for PCC, we take the mean agreement of the three annotations as the ground truth. Note also that neither of the above two metrics requires a hard threshold to determine the important words for evaluation. Whereas R-Prec considers the words with the top-R highest predicted probabilities, PCC directly leverages the predicted probabilities of words as the importance of words for calculation.

#	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**

Table 3: Highlighting performance

5.3 Compared Methods

Zero-shot We fine-tuned the BERT-base model on the e-SNLI$_c$ training set (see Table 2(b)) with the binary token classification cross-entropy objective (See Section 3.3 for details) and used this as a zero-shot approach for 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 2(a)) with the pseudo-label tokens (see pseudo-labeling introduced in Section 3.4) and use this as a pseudo few-shot approach.

Domain-adaptive Using the zero-shot highlighting model as the initialization, we further performed in-domain fine-tuning (see stage S_2^+ in Section 3.4) for domain adaptation.

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 pseudo and soft labeling, respectively, and '$*$' denotes statistical significance with respect to the performance of zero-shot learning (#1) under a paired t-test with $p < 0.05$.

We first focus on the results of the main task on FINAL, where the listed results are those evaluated on the union of the two evaluation sets (including 400 segment pairs in total). As shown in the table, the proposed domain-adaptive approach using 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 performance increase from condition #2 to #3, we observe that warm-up fine-tuning (W.U.) plays an essential role in financial signal highlighting. Similarly, soft la-

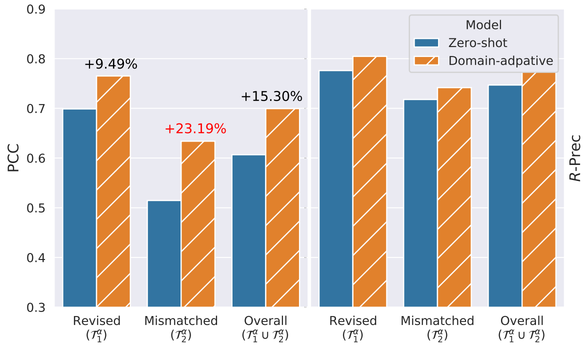

Figure 2: Highlighting effectiveness for different relations, including $\mathcal{T}_1^\alpha$ and $\mathcal{T}_2^\alpha$ with two evaluation metrics

beling is also beneficial for our task, bringing a 10% performance improvement in both evaluation metrics (by comparing the results of conditions #3 and #4). However, from the results of conditions #1 and #3, we observe that adopting pseudo-labeling alone might not be helpful for this task, perhaps because the pseudo-labels constructed by the proposed heuristic approach (see Section 3.3) are too aggressive for unimportant tokens, resulting in a biased highlighting model. In sum, we offer two main observations from Table 3.

- The proposed domain-adaptive fine-tuning with pseudo and soft labeling is effective for signal highlighting in financial reports.
- Warm-up fine-tuning and soft labeling are two crucial components to constructing an effective domain-adaptive highlighting model.

Generalization ability between domains Table 3 also lists the results on the e-SNLI$_c$ testing data: only the model with condition #4 performs on par with or even outperforms that with condition #1 (i.e., zero-shot), showing that the highlighting model fine-tuned by the propose domain-adaptive approach exhibits good generalizability.

5.4.2 Analyses on Different Types of Relationships

To better understand the empirical advantages of the domain-adaptive approach, we further investigate the highlighting performance for different kinds of reference-to-target relations, $\mathcal{T}_1^\alpha$ (*revised*) and $\mathcal{T}_2^\alpha$ (*mismatched*). Figure 2 compares the results of the zero-shot (#1) and domain-adaptive (#4) methods in terms of two metrics. We here focus on PCC, as R-precision considers only the set of important words (i.e., labeled as positive) instead of all the words in each target segment. In the figure, we see that despite the significant PCC improvements on both *revised* and *mismatched* pairs, the benefit of domain adaptation on mismatched pairs is markedly greater than that on revised pairs, yielding a PCC improvement of approximately 23%. Perhaps the important words in the mismatched pairs are more uncertain, necessitating intensive domain adaptation more than those in the revised pairs. Note that we fine-tuned the model on only 30,000 revised segment pairs in $\mathcal{T}_1^\alpha$ for domain adaptation; however, the highlighting results of mismatched pairs $\mathcal{T}_2^\alpha$ exhibit more significant improvement. This suggests that the proposed domain-adaptive approach addresses domain shift and yields a superior ability to infer word importance even for unfamiliar (unseen) relationships (See Appendix E also).

Reference settings		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	$\tilde{r}$	0.5345	0.4582	0.5658	0.4628
Original	r	**0.7865**	**0.7290**	**0.8605**	**0.7566**

Table 4: Impact of referenced knowledge sources

5.5 Ablation Studies

5.5.1 Impact of Referenced Sources

We first determined the impact of the reference segment, which is viewed as the context of a given target segment in terms of discovering the financial signals in the target segment. To this end, for each reference-to-target pair (r, t), we substituted the original reference segment r (i.e., the most syntactically similar segment in the previous years' document $\mathcal{D}_{\ell-1}$) for other text and constructed a few variants of variant-to-target segment pairs for inference using the highlighting model. Specifically, we fixed the target segment but recast the BERT contextualized representation of variant pairs as

- **Empty**: A single `[PAD]` token is used as the reference segment (implying *none* in BERT);
- **Same**: The target segment is used as the reference segment;
- **Random**: A randomly selected segment is used as the reference segment.

In Table 4, the original setup significantly outperforms the other three settings in both FINAL and e-SNLI$_c$, showing that the knowledge provided by the reference segments is critical for capturing important financial signals in the corresponding target segment.

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

Table 5: Different pseudo-labeling approaches

5.5.2 Effect of Lexicon-based labeling

Recall that in Section 3.4, we introduced a heuristic pseudo-labeling approach that views all revised words in target segment t as important words and marks them as positive while we randomly sample other words as negative words. We here test the effect of additionally incorporating an external financial lexicon for pseudo-labeling. Specifically, we adopt the most representative financial sentiment lexicon—the *Loughran–McDonald Master Dictionary* (Loughran and Mcdonald, 2011)—and assume that in addition to the revised words in the heuristic approach, the 3,872 sentiment words in the dictionary also reveal important financial signals (i.e., are labeled as positive). Additionally, we treat the 20K most frequently-occurring words, as well as the standard stopwords, as negative words.

As shown in Table 5, surprisingly, adding the lexicon for pseudo-labeling does not improve performance but instead worsens the highlighting results. Although these financial sentiment words convey important financial signals, they are globally important among all financial reports. However, this characteristic precludes the use of the lexicon for company-specific reference-to-target highlighting, which is focused more on local relationships between a pair of segments.

6 Related Work

Research on financial report analysis has been ongoing for many years, with various studies utilizing both textual and numerical features to identify signals in reports. For instance, some researchers have used the relationship between tokens and quantitative indicators from the financial market to identify financial risks (e.g., Kogan et al., 2009; Tsai and Wang, 2017; Lin et al., 2021). Others have adapted unsupervised methods to recognize information and classify risk factors in financial reports (e.g., Huang and Li, 2011; Lin et al., 2011). However, previous research has mostly focused on risk factors in a global context rather than company-specific signals, which is the focus of this study.

Recently, transformer-based language models such as BERT, GPT-3, and T5 (Devlin et al., 2019; Brown et al., 2020; Raffel et al., 2020) have made significant strides in the summarization task. In 2020, Zmandar et al. (2021a) proposed the Financial Narrative Summarization shared task (FNS 2020), which aims to summarize annual UK reports. While some methods for this task have achieved satisfactory performance using ROUGE as a metric (e.g., Zmandar et al., 2021b; Orzhenovskii, 2021; Gokhan et al., 2021), they have been criticized for sometimes omitting essential signals under a ROUGE-guided policy. Additionally, the signals discovered through these approaches are heavily dependent on high-quality human-annotated summaries, making it challenging to apply them in real-world scenarios.

In the field of NLP, some research has focused on developing rationalizing models related to the concept of our highlighting model. For example, Lei et al. (2016) proposed a method for learning the rationale (words) to justify a model's prediction by selecting a subset of text inputs. More recently, some studies have proposed methods that can rationalize the relationship of sentence pairs, such as natural language inference (Jiang et al., 2021) and query-document relevance (Kim et al., 2022). Additionally, DeYoung et al. (2020) released a benchmark to facilitate the development of interpretable NLP models with faithfulness.

7 Conclusion

This paper addresses the task of identifying rationales as insightful financial signals between two narrative financial reports in consecutive years. We use the reference-to-target structure of financial reports to develop a compare-and-contrast multistage pipeline, comprising mainly of relation recognition and signal highlighting stages. In particular, we propose domain-adaptive learning strategies for signal highlighting, including out-of-domain warm-up and in-domain fine-tuning. Our empirical results confirm the effectiveness of the proposed approaches. We also present the newly constructed FINAL dataset for future research. Future work includes increasing efficiency by integrating dense retrieval methods into our pipeline, improving effectiveness by developing multitask learning on large financial corpora as financial pre-trained representations, and analyzing cross-company relationships beyond year-to-year relationships.

8 Limitations

We identify crucial financial signals in reports which can help financial practitioners to digest long financial documents efficiently. However, factors such as macroeconomics, stock prices, and public policies may affect how a financial practitioner views financial reports in practice. Confidential intelligence or social media may greatly affect the analysis results. Therefore, we limit our task to the scenario in which the content in the reports is the sole information available to users. Accordingly, to prevent bias in the annotation process, we acquire annotations from annotators under similar scenarios (graduate students majoring in accounting or other related fields) rather than from financial professionals.

References

Yash Agrawal, Vivek Anand, S Arunachalam, and Vasudeva Varma. 2021. Hierarchical model for goal guided summarization of annual financial reports. In *Proc. of WWW*, pages 247–254.

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

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. 2006. Analysis of representations for domain adaptation. In *Proc. of NIPS*, pages 137–144.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*, pages 1877–1901.

Chris Buckley and Ellen M. Voorhees. 2000. Evaluating evaluation measure stability. In *Proc. of SIGIR*, page 33–40.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. In *Proc. of NIPS*, pages 9539–9549.

Brian Chivers, Mason P. Jiang, Wonhee Lee, Amy Ng, I. Rapstine, and Natalya Alex Storer. 2022. ANTS: A framework for retrieval of text segments in unstructured documents. In *Proc. of NAACL DeepLo Workshop*, pages 38–47.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proc. of NAACL*, pages 615–621.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proc. of ACL*, pages 4443–4458.

Mine Ertugrul, Jin Lei, Jiaping Qiu, and Chi Wan. 2017. Annual report readability, tone ambiguity, and the cost of borrowing. *J. Financ. Quant. Anal.*, 52(2):811–836.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychol. Bull.*, 76(5):378.

Tuba Gokhan, Phillip Smith, and Mark Lee. 2021. Extractive financial narrative summarisation using sentencebert based clustering. In *Proc. of the FNP Workshop*, pages 94–98.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proc. of ACL*, pages 8342–8360.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proc. of EMNLP-IJCNLP*, pages 4238–4248.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proc. of NIPS*, page 1693–1701.

Ke-Wei Huang and Zhuolun Li. 2011. A multilabel text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Trans. Manag. Inf. Syst.*, 2(3):1–19.

Zhongtao Jiang, Yuanzhe Zhang, Zhao Yang, Jun Zhao, and Kang Liu. 2021. Alignment rationale for natural language inference. In *Proc. of ACL-IJCNLP*, pages 5372–5387.

Youngwoo Kim, Razieh Rahimi, and James Allan. 2022. Alignment rationale for query-document relevance. In *Proc. of SIGIR*, page 2489–2494.

Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proc. of NAACL HLT*, pages 272–280.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proc. of EMNLP*, pages 107–117.

Tian Li, Xiang Chen, Zhen Dong, Kurt Keutzer, and Shanghang Zhang. 2022. Domain-adaptive text classification with structured knowledge from unlabeled data. In *Proc. of IJCAI*, pages 4216–4222.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proc. of ACL*, pages 74–81.

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 Trans. Manag. Inf. Syst.*, 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 *Proc. of ECML-PKDD*, pages 253–268.

Tim Loughran and Bill Mcdonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *J. Finance*, 66(1):35–65.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proc. of EMNLP*, pages 4707–4716.

Mikhail Orzhenovskii. 2021. T5-LONG-EXTRACT at FNS-2021 shared task. In *Proc. of FNP Workshop*, pages 67–69.

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. *arXiv:2210.04722*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP-IJCNLP*, pages 3982–3992.

Navid Rekabsaz, Mihai Lupu, Artem Baklanov, Alexander Dür, Linda Andersson, and Allan Hanbury. 2017. Volatility prediction using financial disclosures sentiments with word embedding-based IR models. In *Proc. of ACL*, pages 1712–1721.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proc. of ACL*, pages 1073–1083.

Gennady Shtekh, Polina Kazakova, Nikita Nikitinsky, and Nikolay Skachkov. 2018. Applying topic segmentation to document-level information retrieval. In *Proc. of CEE-SECR*. Article no. 6.

Ming-Feng Tsai and Chuan-Ju Wang. 2017. On the risk prediction and analysis of soft information in finance reports. *Eur. J. Oper. Res.*, 257(1):243–250.

Haifeng You and Xiao-jun Zhang. 2009. Financial reporting complexity and investor underreaction to 10-K information. *Rev. Account. Stud.*, 14(4):559–586.

Nadhem Zmandar, Mahmoud El-Haj, Paul Rayson, Marina Litvak, Geroge Giannakopoulos, Nikiforos Pittaras, et al. 2021a. The financial narrative summarisation shared task FNS 2021. In *Proc. of FNP Workshop*, pages 120–125.

Nadhem Zmandar, Abhishek Singh, Mahmoud El-Haj, and Paul Rayson. 2021b. Joint abstractive and extractive method for long financial document summarization. In *Proc. of FNP Workshop*, pages 99–105.

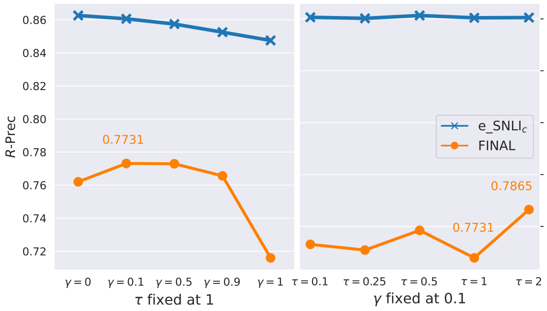

Figure 3: Domain-adaptive labeling

A Training Detail

All our model fine-tuning and inference (in Section 5 and Section 5.5) were conducted on an NVIDIA Tesla V100 32GB GPU. Each model fine-tuning can be done within three hours. We also ran all of the models with shared training settings, including the number of training steps, optimizers, and token batch sizes; we set other related training parameters as the default values.

B Hyperparameter Search

Recall that while the hyperparameter τ in Eq. (2) controls the probability distribution of the word importance, γ in Eq. (3) controls the impact of soft labeling. Figure 3 shows the performance in terms of R-Prec with different hyperparameter settings, where the left panel shows the results of τ fixed at 1 with γ ranging from 0 to 1, and the right panel shows that of γ fixed at 0.1 with τ ranging from 0.1 to 2. In the left panel of the figure, on FINAL, we see that solely adopting cross-entropy loss $\mathcal{L}_{\text{CE}}$ ($\gamma = 1$) is not effective for fine-tuning the signal highlighting model, nor is adopting KL loss $\mathcal{L}_{\text{KL}}$ ($\gamma = 0$) (see Eq. (3)); $\gamma = 0.1$ achieves the best R-Prec. These empirical results again validate the effectiveness of the proposed soft labeling for our highlighting task. In addition, we froze γ at 0.1 and experimented with different settings for the temperature parameter τ, the results of which are shown in the right panel of Figure 3, showing that $\tau = 2$ is the most effective setting. We thus set our final hyperparameters to $\tau = 2$ and $\gamma = 0.1$ to yield the best performance.

C Empirical Thresholds

For the relation recognition procedure in S_1 (see Section 3.2 and Figure 4), we empirically set the thresholds $\epsilon_{\text{syn}} = 0.6296$ and $\epsilon_{\text{sem}} = 0.9011$. Both numbers are the 50 percentiles of the corresponding similarity scores calculated from the reduced segment pair set ($\mathcal{T}$). Note that, in this work, we adopt a rule-based heuristic method for recognizing relations using similarity functions with hard thresholds. We leave the exploration of other similarity functions, thresholds, and approaches to future work.

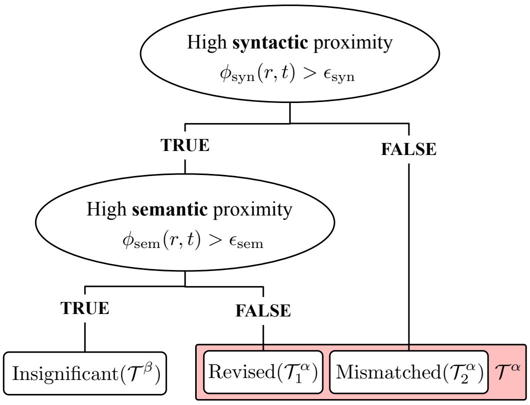
High **syntactic** proximity
$\phi_{\text{syn}}(r,t) > \epsilon_{\text{syn}}$
TRUE
FALSE
High **semantic** proximity
$\phi_{\text{sem}}(r,t) > \epsilon_{\text{sem}}$
TRUE
FALSE
Insignificant($\mathcal{T}^{\beta}$)
Revised($\mathcal{T}_1^{\alpha}$)
Mismatched($\mathcal{T}_2^{\alpha}$)
$\mathcal{T}^{\alpha}$

Figure 4: Relation recognition

D Annotation Guidelines

For each segment pair, the annotators were to focus on the semantic difference regarding the reference-to-target relationship and annotate words in the target segment as positive when the words were considered important financial signals. The following guidelines were given for the annotators' reference.

- Changes: Changing numbers or objects are important signals in financial reports (e.g., sales, cost, partnership, products, etc.).
- Opposition: Descriptive phrases that indicate distant semantic meanings (e.g., increased/decreased, effective/ineffective, etc.).
- Precise: Labeling words with high confidence as positive only (i.e., leaving ambiguous words as negative).
- Extra information: Identifying new information according to the context, for which the annotators considered the reference segment as the context (e.g., new policy, canceled deals, newly published products, etc.).

E Empirical Cases

In Table 6, we take two revised segment pairs ($\mathcal{T}_1^{\alpha}$) and two mismatched segment pairs ($\mathcal{T}_2^{\alpha}$) as examples. The underlined words are with the top-k highest importance predicted by the proposed multistage pipeline.

(a) Empirical examples of the revised segment pair ($\mathcal{T}_1^{\alpha}$) [$k = 5$]

Reference segment Gross margin from manufacturing operations as a percentage of manufacturing revenues increased to 27% for the year ended December 31, 2014, from 23% for the comparable prior year period.

Target segment Gross margin from manufacturing operations as a percentage of manufacturing revenues decreased to 15% for the year ended December 31, 2016 from 23% for the comparable prior year period.

Reference segment We believe the increased sales achieved by our stores are the result of store growth and the high levels of customer service provided by our well-trained and technically proficient Team Members, superior inventory availability, including same day and over-night access to inventory in our regional distribution centers, enhanced services and programs offered in our stores, a broader selection of product offerings in most stores with a dynamic catalog system to identify and source parts, a targeted promotional and advertising effort through a variety of media and localized promotional events, continued improvement in the merchandising and store layouts of our stores, compensation programs for all store Team Members that provide incentives for performance and our continued focus on serving both DIY and professional service provider customers.

Target segment We believe the increased sales achieved by our stores were the result of store growth, sales from one additional day due to Leap Day for the year ended December 31, 2016, sales from the acquired 48 Bond stores, the high levels of customer service provided by our well-trained and technically proficient Team Members, superior inventory availability, including same day and over-night access to inventory in our regional distribution centers, enhanced services and programs offered in our stores, a broader selection of product offerings in most stores with a dynamic catalog system to identify and source parts, a targeted promotional and advertising effort through a variety of media and localized promotional events, continued improvement in the merchandising and store layouts of our stores, compensation programs for all store Team Members that provide incentives for performance and our continued focus on serving both DIY and professional service provider customers.

(b) Empirical examples of the mismatched segment pair ($\mathcal{T}_2^{\alpha}$) [$k = 10$]

Reference segment This increase of 1.0%, as a percentage of revenues, was primarily attributable to higher compensation costs of 0.4% primarily related to higher wage rates, higher facility-related costs of 0.2% principally from the expansion of U.S. facilities and lease termination costs in connection with the Fourth Quarter 2011 Exit Plan, higher software maintenance of 0.2%, higher legal and professional fees of 0.1%, higher taxes of 0.1% and higher other costs of 0.3%, partially offset by lower equipment and maintenance costs of 0.3%.

Target segment The decrease in Americas general and administrative expenses, as a percentage of revenues, was primarily attributable to lower compensation costs of 0.6%, lower facility-related costs of 0.4% due to rationalization of facilities, lower equipment and maintenance costs of 0.2% and lower other costs of 0.1%.

Reference segment The remaining capacity is expected to be placed into service in line with the expected in-service date of the Sandpiper Project.

Target segment Three external parties filed motions requesting that the scoping process be re-opened or that a comment period be established because of the issuance of the Consent Decree settling the Line 6B pipeline crude oil release in Marshall, Michigan and the withdrawal of regulatory applications pending with the MNPUC with respect to the Sandpiper Project discussed above.

Table 6: Empirical cases in the FINAL evaluation set