Semi-supervised News Discourse Profiling with Contrastive Learning

Ming Li
Texas A&M University
liming@tamu.edu

Ruihong Huang
Texas A&M University
huangrh@cse.tamu.edu

Abstract

News Discourse Profiling seeks to scrutinize the event-related role of each sentence in a news article and has been proven useful across various downstream applications. Specifically, within the context of a given news discourse, each sentence is assigned to a pre-defined category contingent upon its depiction of the news event structure. However, existing approaches suffer from an inadequacy of available human-annotated data, due to the laborious and time-intensive nature of generating discourse-level annotations. In this paper, we present a novel approach, denoted as Intra-document Contrastive Learning with Distillation (ICLD), for addressing the news discourse profiling task, capitalizing on its unique structural characteristics. Notably, we are the first to apply a semi-supervised methodology within this task paradigm, and evaluation demonstrates the effectiveness of the presented approach. Codes, models, and data will be available. [1]

1 Introduction

News discourse profiling (Choubey et al., 2020) is a specialized task aimed at comprehensively analyzing the structural aspects of news articles and effectively categorizing each sentence based on its contextual depiction of news events. Therefore, this is a document-level task with sentence-level predictions (Li et al., 2022a), which has been proven useful in several downstream tasks, including text simplification (Zhang et al., 2022a), media bias analysis (Lei et al., 2022), event coreference resolution (Choubey et al., 2020), RST-style Discourse Parsing (Li and Huang, 2023) and temporal dependency graph building (Choubey and Huang, 2022).

Nevertheless, as a discourse-level task, the process of creating annotations entails a substantial investment of time and labor. The absence of human-annotated data poses a significant obstacle to the practical implementation of news discourse profiling, despite the relatively straightforward acquisition of unlabeled news articles. Building upon the aforementioned rationale, our impetus resides in introducing more unlabeled data for training purposes and formulating a semi-supervised methodology tailored to this particular task structure.

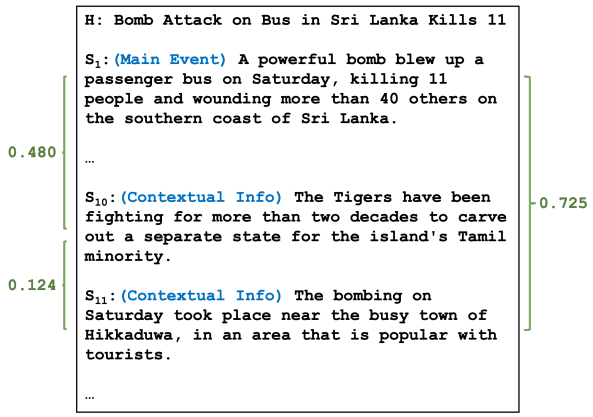

Figure 1: An example of news articles in the human-annotated data. H represents the headline of the news and S_i represents the ith sentence in the news. In parenthesis (blue) are news discourse profiling labels assigned to the respective sentences. On the left and right sides (green), we provide the cosine similarity values derived from the sentence embeddings generated by Google's Universal Sentence Encoder. In this example, S_1 and S_{11} are semantically similar (0.725) but have different labels. However, S_{10} and S_{11} are in the same category even though they are not semantically similar (0.124).

Contrastive learning (Zhang et al., 2022b; Le-Khac et al., 2020; Albelwi, 2022) can effectively make use of unlabeled data and has been developed greatly recently which aims to learn effective representations of words, sentences, or discourses by pulling semantically close samples together and pushing away others (Gao et al., 2021). A fundamental premise underlying contrastive learning is that the features acquired by encoders, via self-identification, encompass crucial information capable of not only distinguishing individual in-

[1] https://github.com/MingLiiii/ICLD

stances but also discerning disparities across different classes (van den Oord et al., 2019). However, in news discourse profiling, the classification of each sentence is more profoundly influenced by the collective discourse structure and its interrelationships with other sentences, rather than relying solely on the inherent semantic meanings of individual sentences, which poses difficulties in designing a contrastive learning methodology for this task.

As depicted in Figure 1, we present an illustrative instance selected from human-annotated datasets. H denotes the news headline, while S_i represents the i-th sentence within the news[2]. In parenthesis (blue) are news discourse profiling labels assigned to the respective sentences. On the left and right sides (green), we provide the cosine similarity values derived from the sentence embeddings generated by Google's Universal Sentence Encoder (Yang et al., 2020).

Although S_{10} and S_{11} serve similar functions by elucidating the terrorists' workplace and specifying the precise detonation site of the bomb, their similarity score is a mere 0.124. Conversely, S_1 and S_{11} exhibit a higher similarity score of 0.725, despite their distinct narrative roles in conveying this news account. Consequently, it is evident that a substantial discrepancy exists between the assigned sentence categories and their underlying semantic significance, underscoring a pronounced misalignment. In such a scenario, conventional sentence-level contrastive learning approaches prove inadequate for enhancing news discourse profiling, primarily due to their emphasis on capturing sentence-level semantic meanings. Furthermore, standalone sentences devoid of contextual information lack the capacity to effectively represent the intricate high-level event structures characterizing the entire discourse.

Building upon the aforementioned discussions, our objective is to establish an embedding space that not only captures semantic similarities but also incorporates the underlying event structure. Diverging from conventional contrastive learning methods that construct instance pairs through self-supervision, our approach operates in a semi-supervised manner. Thus we present a novel semi-supervised approach for news discourse profiling, termed Intra-document Contrastive Learning with Distillation (ICLD), specifically designed for this discourse-level task. In our proposed method, we employ a teacher model to predict silver labels for unlabeled news articles that have not been previously seen. These predicted labels act as guiding signals for the construction of positive and negative sentence pairs within each document, facilitating the contrastive learning process. Furthermore, our method incorporates intra-document contrastive learning along with an additional knowledge distillation component. This serves two purposes: firstly, to ensure the interaction between the target sentence and its contextual surroundings, and secondly, to further prevent the collapse of the contrastive aspect into simply learning the semantics similarities of individual sentences.

Extensive experimental evaluations have been conducted, confirming the efficacy of our proposed method. By incorporating a larger volume of readily accessible unlabeled news articles, we achieve a significant improvement in news discourse profiling performance. Notably, to the best of our knowledge, we are the first to address this particular task structure and propose a semi-supervised methodology to tackle it.

2 Related Work

2.1 Contrastive Learning

Recently the technique contrastive learning has been widely used in unsupervised and self-supervised learning, which greatly improved the performance of both visual and language representation (Ting Chen and Hinton, 2020; Kaiming He et al., 2020; Tianyu Gao, 2021; Wu et al., 2020; Zhang et al., 2021; Janson et al., 2021). It learns the data representation by pushing away negative samples and pulling close the positive samples where InfoNCE (van den Oord et al., 2019) objective is mostly used. Ideally, this would update the encoder to carry enough information for both sample identification and downstream classification.

After achieving great success in computer vision tasks (Ting Chen and Hinton, 2020; Kaiming He et al., 2020), contrastive learning methods are then applied to the Natural language processing (NLP) area for sentence representation learning (Tianyu Gao, 2021; Wu et al., 2020; Zhang et al., 2021; Janson et al., 2021). One of the main methodological differences among these works is the method of data augmentation to generate positive pairs. CLEAR (Wu et al., 2020) utilizes word

[2]The ordering of sentences is important in analyzing the event structure of news articles.

deletion, span deletion, reordering, and substitution for data augmentation. It calculates objectives on both the token level and the sentence level. DeCLUTR (Giorgi et al., 2020) treats sentences from the same documents as positive pairs, while sentences from different documents as negative pairs. SimSCE (Tianyu Gao, 2021) utilizes the dropout in the pretrained word encoder and is proven to be an efficient way of augmentation. Leveraging the foundational concepts of SimCSE, a plethora of subsequent research endeavors have sought to enhance this framework through the incorporation of advanced auxiliary training objectives (Chuang et al., 2022; Nishikawa et al., 2022; Zhou et al., 2023; Wu et al., 2022), and (Chanchani and Huang, 2023) recently proposes maximizing alignment between texts and composition of their phrasal constituents.

2.2 Knowledge Distillation

Knowledge distillation is first proposed by (Hinton et al., 2015) for model compression by minimizing the KL divergence between the output distributions of the teacher model and the student model.

In NLP tasks, large pretrained language models have achieved remarkable performance (Devlin et al., 2019; Qiu et al., 2020; AlKhamissi et al., 2022). Knowledge distillation is one way to retain comparable performance as large models with relatively compact models. (Yimeng Wu et al., 2021) effectively compresses models like BERT-base to BERT-4. (Dongha Choi and Lee, 2022) comes up with a framework for finetuning a domain-specific pretrained large language model as a teacher, then uses activation boundary distillation to teach domain knowledge to another language model. (Liu et al., 2022) compares the effect of knowledge in three different levels: token level, span level, and sample level among which the sample level maintains most of the knowledge. (Huang et al., 2022) pretrains and finetunes a teacher model without pruning, then progressively replaces layers of the teacher model with the student model learned by knowledge distillation, which mitigates the overfitting in finetuning pretrained language models.

3 The Semi-supervised Method

Our semi-supervised approach consists of two learning phases, the first phase of Intra-document Contrastive Learning with Distillation (ICLD) exclusively utilizes unlabeled news articles and the second phase brings back human annotations to better calibrate the model.

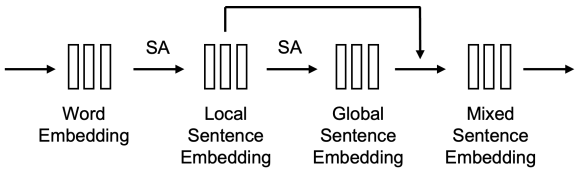

Figure 2: A brief illustration of our model structure following Li et al. (2022b). *SA* represents the self-attention module. Upon receiving an input discourse, initial word embeddings are generated using a language model. Subsequently, two self-attention modules are employed to obtain both local and global sentence embeddings. The mixed sentence embeddings are derived by adding the local and global sentence embeddings, followed by the integration of a fully connected layer. Finally, a classification layer is applied to yield the final prediction. The intra-document contrastive learning phase will be implemented upon mixed sentence embeddings.

3.1 Model Structure

Within the context of the semi-supervised framework, the teacher model encounters a substantial volume of unlabeled news articles that may exhibit diverse distributions distinct from the human-annotated training set. Consequently, to offer more reliable guidance, the teacher model must possess strong generalization capabilities, ensuring the accuracy of the generated labels for unseen data instances. In light of these considerations, we opt for LiMNet (Li et al., 2022b), incorporating the robust T5 large language model (Raffel et al., 2020), as our selected teacher model due to its commendable performance and generalization abilities.

Our student models adopt the same structural framework as LiMNet, leveraging small language models such as Longformer (Beltagy et al., 2020) as the default choice. During training, the weights of these student models are iteratively updated. It is worth noting that the teacher model is solely trained using the original annotated data, adhering to the identical configuration outlined in its original paper. Subsequently, the automatically collected unlabeled news articles are fed into this well-trained teacher model, enabling the derivation of probability distributions across different sentence categories.

Figure 2 illustrates the simplified model architecture of LiMNet, where two self-attention modules (Bahdanau et al., 2014; Chorowski et al., 2015) are utilized. The first self-attention module focuses

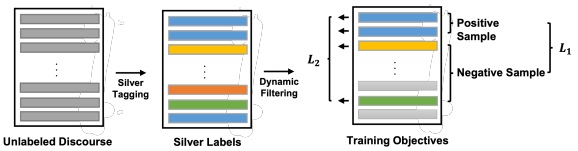

Figure 3: An overview of our proposed method. For a given unlabeled discourse, the teacher model is first utilized to tag the sentences of the discourse to obtain the silver labels, where different colors denote different classes. Then dynamic filtering is implemented to filter those sentences with relatively low confidence, the gray color means the sentences are neglected. Then, sentences in the same category are randomly sampled as positive samples while sentences with different categories are randomly sampled as negative ones, on which is the contrastive training objective L_1 calculated. In the meantime, an extra distillation phase L_2 is implemented on unfiltered sentences to avoid the collapse solution.

on capturing interactions among word embeddings within the given sentence, thereby producing a localized representation as the local sentence embedding. The second self-attention module leverages the interaction between the specific sentence embedding and the contextual word embeddings to derive a comprehensive global sentence embedding. Finally, a fully connected layer is employed for the purpose of final classification.

3.2 Phase One: Intra-document Contrastive Learning with Distillation (ICLD)

During the ICLD phase, exclusively unlabeled news articles are utilized, figure 3 illustrates the workflow of intra-document contrastive Learning with distillation (ICLD).

Random Sentence Filtering Due to the potential distribution shift between the human-annotated data and the newly acquired unlabeled data, strictly adhering to the generated silver labels may lead to suboptimal outcomes. Moreover, the imperfections of the teacher model could be transmitted to the student models through the inclusion of noisy silver labels. To address this challenge, we propose a solution involving the random filtering of sentences exhibiting relatively lower confidence.

Considering the variability in confidence distributions across different categories, it is impractical to manually establish a specific static threshold for each category. Instead, we utilize a more flexible approach that leverages the intrinsic characteristics of the teacher models and the newly collected articles. Specifically, we adopt the k-th percentile approach for determining flexible thresholds, based on the silver label confidences estimated by the teacher model. These thresholds are dynamically adjusted as the teacher model or unseen data undergo modifications. Consequently, sentences with confidences lower than their respective thresholds are subjected to filtering with a probability of 0.5 during each epoch. This stochastic filtering approach is employed due to the inherent uncertainty associated with low-confidence sentences, as their definitive correctness cannot be ascertained in the absence of golden labels. By employing adaptable thresholds, we mitigate the reliance on predefined confidence thresholds for individual categories, allowing for a more tailored and nuanced filtering process.

Intra-document Contrastive Learning (ICL)

The generation of positive pairs for contrastive learning follows a specific procedure: under the supervision of silver labels, unfiltered sentence pairs of the same category within each document are randomly sampled without replacement until no sentences remain belonging to the same category. Then, negative pairs are randomly sampled without replacement, from the remaining sentences where no sentence pair shares the same silver label until no additional pairs can be generated. Specifically, once the silver labels are procured, for each label with more than two associated sentences, we randomly select two sentences. This process continues until just one or no sentence remains for that label, constituting our positive samples. For the remaining sentences, we repeat a similar sampling of two sentences at a time to form negative pairs until only one sentence or none remains in the document. This sampling process guarantees diverse positive and negative pairs, facilitating effective contrastive learning within the intra-document context.

Following the establishment of positive and negative pairs, the contrastive learning constraint is imposed on the mixed sentence embeddings derived from the input discourse. Notably, the mixed sentence embeddings encompass crucial contextual information essential for comprehending the news event content, distinguishing them from lower-level sentence embeddings. Thus, the contrastive learning process focuses on leveraging contextual embeddings to enhance the discriminative ability and semantic understanding of the news event representations. Cosine similarity is used as the measurement of similarity between two embeddings, and

the contrastive learning constraint is formulated as:

$$L_1 = -log\frac{m \cdot \sum_{i=1}^{n} e^{cos(g_i, g_i^-)}/\tau}{n \cdot \sum_{j=1}^{m} e^{cos(g_j, g_j^+)}/\tau + \beta} \quad (1)$$

where g_i and g_i^- represent global sentence embeddings in negative pair, g_j and g_j^+ represent sentence embeddings in positive pair. n is the number of positive pairs and m is the number of negative pairs in this discourse. τ represents the temperature rate which is in the range of 0 to 1. β represents a small value to avoid a division by zero and is set to $1e-6$ by default.

Knowledge Distillation

In contrast to tasks that primarily emphasize the semantic meanings of individual sentences, news discourse profiling directs its attention toward the collective representation of sentences in describing news events. To avoid the potential collapse of task-specific intra-document contrastive learning towards standard contrastive learning, which prioritizes the semantic meanings of individual sentences, the incorporation of explicit guidance and simultaneous distillation becomes imperative.

To address this challenge, silver labels simultaneously serve as direct training guidance for the student model. This facilitates a more explicit and informed learning process. Concurrently, flexible threshold-based random filtering is applied to eliminate low-confidence sentences, ensuring the optimization of the student model with reliable and informative training instances. By leveraging these measures, this task-specific intra-document contrastive learning remains focused on capturing the interdependencies and contextual cues within the news discourse, promoting a more accurate and contextually rich news discourse profiling.

Different from training with human-annotated data where cross-entropy is used, we choose to minimize the mean square error (MSE) of probability distribution between the silver labels and the student model, which can be formulated as:

$$L_2 = \frac{1}{l}\sum_{i=1}^{l}(y_i - \hat{y}_i)^2 \quad (2)$$

where $\hat{y}_i$ represents the probability distribution from the student model, and y_i represents the silver labels generated by the teacher model. l is the number of unfiltered sentences.

Contrastive learning projects sentence embeddings to the desired space while knowledge distillation aims to build mappings between sentence representations and discourse role classes, therefore, balancing these two objectives can slow down classifier formation. Empirically, we found that it facilitates training to enable both contrastive learning and knowledge distillation for the first few training epochs and then continues to train for more epochs with knowledge distillation as the only learning objective. When both intra-document contrastive Learning and knowledge distillation are enabled, the overall learning objective is the summation of L_1 and L_2.

3.3 Phase Two: Final Finetuning

After training on unlabeled data in the ICLD phase, the human-annotated golden data are utilized to better calibrate the model. The final model finetuning will use cross-entropy loss as the objective:

$$L_3 = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad (3)$$

where $\hat{y}$ represents the probability distribution from the student model, and y represents the human-annotated gold labels. C represents the number of classes. Overall, L_1 and L_2 are utilized upon unlabeled data, and L_3 is utilized upon human-annotated data.

4 Evaluation

4.1 Dataset

Labeled data

The NewsDiscourse dataset (Choubey et al., 2020) we use is designed for the task of News Discourse Profiling, which consists of 802 news articles (18, 155 sentences). These news discourses are sampled from three news sources including NYT, Xinhua and Reuters and they are in four domains including business, crime, disaster, and politics. Each sentence in this corpus is labeled with one of eight content types[3] representing what role it plays in reporting a news story or the "None" class, following

[3]The eight content types are grouped into three categories: Main Content, Context-informing Content, and Additional Supportive Content. In Main Content, there are two fine-grained categories: *Main Event* which introduces the most important event relating to the major subjects of news discourse, and *Consequence* which represents content that is triggered by the main news event. In Context-informing Content, there are two fine-grained categories: *Previous Event* which precedes the the main event and now acts as possible causes or preconditions for the main event, and *Current Context* which covers all the context informing the main event. In Additional Supportive Content, there are four fine-grained cat-

the news content schemata proposed by Van Dijk (Van Dijk, 1985, 1988). Following Choubey and Huang (2021), we use 502 documents for training, 100 documents for validation, and 200 documents for testing. All the models are evaluated by calculating the micro F1 and macro Precision, Recall, and F1 scores are implemented form the scikit-learn library (Pedregosa et al., 2011).

Unlabeled data

To simulate the real-world application scenario where the data distribution can largely vary from the existing human-annotated data, we deliberately avoid following the same source and domains from the existing data. Unlabeled news articles are collected from CNN with a variety of domains including business, entertainment, health, politics, sports, style, travel, and world. These unlabeled data contain $10,337$ news articles with $135,057$ sentences and all of these data will be used in our method. The improvement made by our method using these data proves that our method is effective even when the unlabeled data have different distributions than the original data.

4.2 Implementation Details

All experiments are implemented in the PyTorch platform (Paszke et al., 2019) with one NVIDIA A100 graphic card.

All pre-trained language models used in this paper are implemented from *huggingface* (Wolf et al., 2019). Our teacher model utilizes *large* version of T5 as the pretrained language model, while all our student models use the *base* version of pretrained language models with the output dimension of 768. The parameters of the language model in the teacher model are fixed all the time including its training phase, and the parameters of the language model in the student models are not fixed and are updated during training.

Our models are trained using Adam optimizer (Kingma and Ba, 2014) with the hyper-parameters betas=[0.9, 0.999], eps=$1e-8$ and the learning rate is set to $5e-6$ for 25 epochs. The dropout rate (Srivastava et al., 2014) is set to 0.5. Intra-document contrastive learning with distillation (ICLD) is applied in the first 3 epochs, where both L_1 and L_2 are utilized. Knowledge distillation from unlabeled data will continue for another 12 epochs with only L_2 as the learning objective. In the first 15 epochs, only unlabeled data with silver labels are utilized. Then, we will continue to train for 10 epochs using only the original well-labeled data and the learning objective of L_3. We use the 50 percentile as the threshold and filter sentences with the probability of 0.5 by default. τ in contrastive learning constraint is set to 1 by default.

egories: *Historical Event* which represents events that precede the main event in months or years, *Anecdotal Event* which represents unverified events of a person or situation, *Evaluation* which represents opinionated contents including reactions from immediate participants, experts, known personalities, as well as journalists or news sources and *Expectation* represents speculations and projected consequences.

4.3 Ablation Study

	Macro			Micro
	Precision	Recall	F1	F1
The full model	**67.9**	**68.8**	**68.3**	**71.0**
w/o ICL	66.9	68.3	67.5	70.1
w/o Distillation	63.2	65.0	63.6	66.5
Positive Only	67.0	68.0	67.4	70.5
Negative Only	67.6	69.2	**68.3**	70.8

Table 1: Ablation experiments of our ICLD method. *w/o ICL* represents the model where intra-document contrastive learning is not utilized, which becomes a standard knowledge distillation method. *w/o Distillation* represents the model where no extra distillation is utilized simultaneously with intra-document contrastive learning. *Positive Only* and *Negative Only* represent experiments where only positive or negative pairs are sampled and calculated in the contrastive constraint.

To evaluate the individual contributions of different components in our proposed Intra-document Contrastive Learning with Distillation approach, an ablation study is conducted, and the results are presented in Table 1. The model labeled as *Without ICL* refers to the variant where intra-document contrastive learning is not incorporated. In this configuration, the model is trained solely with silver and real labels using the same configuration as our final ICLD model, which essentially functions as a standard knowledge distillation method. Comparing the performance of this variant with the complete ICLD model, we observe improvements across all metrics, demonstrating the efficacy of our intra-document contrastive learning component.

On the other hand, the model denoted as *Without Distillation* represents the variant where no distillation process is executed concurrently with intra-document contrastive learning. Notably, this configuration exhibits a significant decline in performance. The absence of the distillation component leads to the collapse of intra-document contrastive learning into standard contrastive learning, where the classification of instance pairs primar-

	Macro			Micro
	Precision	Recall	F1	F1
Probability of 0	67.2	68.3	67.6	70.5
Probability of 0.3	67.4	68.4	67.8	70.6
Probability of 0.5	**67.9**	68.8	**68.3**	**71.0**
Probability of 0.8	67.8	**68.9**	68.2	70.8
Probability of 1.0	67.8	68.3	67.9	**71.0**

Table 2: Extensive experiments with respect to the effects of random filtering. The probability value determines the likelihood of filtering sentences based on their confidence scores. When the probability is set to 0, no filtering is applied, while a probability of 1.0 indicates that every sentence with a confidence score below the threshold will be filtered.

ily relies on their semantic meanings rather than task-specific discourse structure information. This comparative analysis underscores the indispensable nature of the additional distillation process, which furnishes the necessary guidance for the model to acquire task-specific structural information. Overall, the ablation study highlights the importance of both intra-document contrastive learning and distillation in our approach, as they collectively contribute to the enhanced performance in capturing the intricate structure and nuances of news discourse profiling.

In addition, we present experimental results with respect to the sentence pair selection strategy. *Sample Positive* and *Sample Negative* represent experiments where only positive or negative pairs are sampled and calculated in the contrastive constraint. Comparing the performance of these variants, we find that using only positive pairs yields inferior results compared to using only negative pairs. This observation is reasonable as relying solely on positive pairs fails to establish a clear decision boundary and the negative samples play the dominant role in this contrastive learning phase.

4.4 Effects of Random Filtering

In this section, we conducted experiments to investigate the effect of different random filtering probabilities ranging from 0 to 1.0, as presented in Table 2. The default filtering threshold for these models was set to the 50th percentile[4]. When the filtering probability is set to 0, no sentences are filtered out. However, since there are no golden labels available for the newly collected news articles, the accuracy of the generated silver labels cannot be guaranteed. Without random filtering, the model might learn from potential noise in the silver labels, resulting in the lowest performance observed in the *Probability of* 0 experiment. On the other hand, when the filtering probability is set to 1.0, sentences with confidence scores lower than the threshold are entirely filtered out. However, filtering with a high probability is not optimal as it cannot be asserted that the low-confidence instances are definitively incorrect. Filtering out all of these low-confidence samples directly eliminates the possibility for the model to learn from them. Therefore, we chose a filtering probability of 0.5 as the default setting.

[4]The impact of using different confidence thresholds is discussed in Appendix A.

	Macro			Micro
	Precision	Recall	F1	F1
Longformer (baseline)	66.2	62.3	63.4	68.7
Longformer (ICLD)	**67.9**	**68.8**	**68.3**	**71.0**
RoBERTa (baseline)	66.6	61.5	62.9	69.0
RoBERTa (ICLD)	67.2	67.1	67.1	70.8

Table 3: Comparisons with baseline models. Baseline models are trained with human-annotated data only.

4.5 Comparison with Baselines

In this section, we compare our ICLD model with baselines using only original human-annotated data. *Longformer (baseline)* and *Longformer (ICLD)* utilize pretrained Longformer (Beltagy et al., 2020) (base version) implemented from *huggingface* (Wolf et al., 2019). Compared with the baseline model where only human-annotated data is utilized, our ICLD model improves the macro F1 score by 4.9 percent and micro F1 score by 2.3 percent. *RoBERTa (baseline)* and *RoBERTa (ICLD)* utilize pretrained RoBERTa (Liu et al., 2019) (base version) implemented from *huggingface* (Wolf et al., 2019). Compared with the baseline model where only human-annotated data is utilized, our ICLD model improves the macro F1 score by 4.2 percent and micro F1 score by 1.8 percent. These comparisons verify the effectiveness of our proposed method, which improves news discourse profiling by a large margin.

Furthermore, it is observed that models utilizing Longformer demonstrate superior performance compared to those utilizing RoBERTa. RoBERTa is specifically designed to handle inputs within a token limit of 512, whereas news articles often exceed this limit. Consequently, we partition lengthy articles into multiple segments before feeding them into RoBERTa. In this process, sentences belonging to different segments are unable to interact with each other, resulting in the absence of comprehen-

	Macro			Micro
	Precision	Recall	F1	F1
(Choubey et al., 2020)	56.9	53.7	54.4	60.9
(Choubey and Huang, 2021)	58.7	56.4	57.0	62.2
(Spangher et al., 2021)	–	–	63.5	67.5
(Li et al., 2022b)	**68.2**	63.9	65.6	69.7
Our ICLD model	67.9	**68.8**	**68.3**	**71.0**

Table 4: Comparison with previous methods. Spangher et al. (2021) does not report their macro precision and recall scores.

sive global contextual information within their corresponding sentence embeddings. Consequently, the performance is adversely affected. On the contrary, Longformer is explicitly designed to handle long inputs, thereby circumventing this potential segmentation issue. All sentences within a discourse can be collectively modeled, resulting in enhanced performance.

Notably, the performance of these two baseline models is relatively comparable since the standard training process does not fully exploit contextual information. For instance, when predicting the category of the first sentence, it is likely that knowledge of the last sentence is unexploited. Consequently, document splitting has minimal impact on performance. However, in our proposed contrastive-based method, the first and last sentences may be randomly selected in one positive or negative pair. In such cases, the sentence embeddings of these two sentences need to be compared and updated directly. With Longformer, all sentences share a similar contextual environment and possess a holistic view of the entire document, thereby justifying the comparison and update process. However, if the document is split when employing RoBERTa, these two sentences might belong to distinct semantic environments, making the comparison unjustifiable.

4.6 Comparison with Previous Methods

Table 4 shows the performances of previous methods. In contrast to previous approaches, our semi-supervised method leverages newly introduced unlabeled data, resulting in a substantial improvement in news discourse profiling performance. Choubey and Huang (2021) utilizes sub-topic information to guide the embedding extraction in an actor-critic manner. Spangher et al. (2021) improves news discourse profiling performance by utilizing the multitask training from several discourse datasets.[5] Li et al. (2022b) mainly focuses on alleviating overfitting for discourse-level tasks and the performance in the table is based on T5 language model (Raffel et al., 2020), which serves as our teacher model. It is worth noting that our ICLD model with the *base* versions of Longformer or RoBERTa surpasses the performance of Li et al. (2022b) where the *large* version of T5 language model is utilized. Considering the huge discrepancy in model sizes and the training corpus utilized for these language models, we assert that our proposed method exhibits a significant capability for this task.

5 Conclusion

In this work, we introduce the Intra-document Contrastive Learning with Distillation (ICLD) method for news discourse profiling, which leverages unlabeled data to enhance performance. News discourse profiling is a discourse-level task where the prediction of each sentence is intricately linked to the overall event structure of the entire discourse, rather than solely relying on the semantic meaning of individual sentences. To the best of our knowledge, we are the first to address this unique task structure and propose a semi-supervised approach to tackle it effectively. The dataset we collected emulates real-world scenarios, encompassing potential domain shifts, and our method demonstrates robust performance in such settings, affirming the efficacy of our proposed approach.

Limitations

In this task, it is observed that sentences with similar semantic meanings can be assigned to different categories. Therefore, our objective is to establish an embedding space that not only captures semantic similarities but also incorporates the underlying event structure. To achieve this, we designed a contrastive learning based approach. However, it is important to note that our contrastive learning method is not necessarily the optimal solution when compared to other possible semi-supervised methods. The primary motivation of this paper is to address the unique task structure and propose a semi-supervised method that is specifically designed for this task. We do not claim that our method is comprehensive or superior but only serves as an initial exploration of a semi-supervised approach tailored to this intriguing task structure.

[5]For the detailed description of datasets, please see the Appendix in Spangher et al. (2021).

Acknowledgements

We thank the anonymous reviewers for their valuable feedback and input. We gratefully acknowledge support from the National Science Foundation (NSF) via the awards IIS-1942918 and IIS-2127746. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing.

References

Saleh Albelwi. 2022. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4):551.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Sachin Chanchani and Ruihong Huang. 2023. Composition-contrastive learning for sentence embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15836–15848, Toronto, Canada. Association for Computational Linguistics.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling news discourse structure using explicit subtopic structures guided critics. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1594–1605, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2022. Modeling document-level temporal structures for building temporal dependency graphs. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 357–365, Online only. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

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.

HongSeok Choi Dongha Choi and Hyunju Lee. 2022. Domain knowledge transferring for pre-trained language model via calibrated activation boundary distillation. *Annual Meeting of the Association for Computational Linguistics*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2020. Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv:1503.02531*.

Shaoyi Huang, Dongkuan Xu, Ian Yen, Yijue Wang, Sung-En Chang, Bingbing Li, Shiyang Chen, Mimi Xie, Sanguthevar Rajasekaran, Hang Liu, and Caiwen Ding. 2022. Sparse progressive distillation: Resolving overfitting under pretrain-and-finetune paradigm. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 190–200, Dublin, Ireland. Association for Computational Linguistics.

Sverker Janson, Evangelina Gogoulou, Erik Ylipää, Amaru Cuba Gyllensten, and Magnus Sahlgren. 2021.

Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations, 2021*.

Yuxin Wu Kaiming He, Haoqi Fan, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Yuanyuan Lei, Ruihong Huang, Lu Wang, and Nick Beauchamp. 2022. Sentence-level media bias analysis informed by discourse structures. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10040–10050.

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022a. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):1–12.

Ming Li and Ruihong Huang. 2023. Rst-style discourse parsing guided by document-level content structures. *arXiv preprint arXiv:2309.04141*.

Ming Li, Sijing Yu, and Ruihong Huang. 2022b. Less is more: Simplifying feature extractors prevents overfitting for neural discourse parsing models. *arXiv preprint arXiv:2210.09537*.

Chang Liu, Chongyang Tao, Jiazhan Feng, and Dongyan Zhao. 2022. Multi-granularity structural knowledge distillation for language model compression. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1001–1011, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Alexander Spangher, Jonathan May, Sz-Rung Shiang, and Lingjia Deng. 2021. Multitask semi-supervised learning for class-imbalanced discourse classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 498–517, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Danqi Chen Tianyu Gao, Xingcheng Yao. 2021. Simcse: Simple contrastive learning of sentence embeddings. *EMNLP*.

Mohammad Norouzi Ting Chen, Simon Kornblith and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. *International conference on machine learning*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

Teun A Van Dijk. 1985. Structures of news in the press. *Discourse and communication: New approaches to the analysis of mass media discourse and communication*, 10:69.

Teun A Van Dijk. 1988. News analysis. *Case Studies of International and National News in the Press. New Jersey: Lawrence*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xing Wu, Chaochen Gao, Zijia Lin, Jizhong Han, Zhongyuan Wang, and Songlin Hu. 2022. InfoCSE:

Information-aggregated contrastive learning of sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3060–3070, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020. Multilingual universal sentence encoder for semantic retrieval. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.

Abbas Ghaddar Yimeng Wu, Mehdi Rezagholizadeh, Md Akmal Haidar, and Ali Ghodsi. 2021. Universal-kd: Attention-based output-grounded intermediate layer knowledge distillation. *Conference on Empirical Methods in Natural Language Processing*.

Bohan Zhang, Prafulla Kumar Choubey, and Ruihong Huang. 2022a. Predicting sentence deletions for text simplification using a functional discourse structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 255–261, Dublin, Ireland. Association for Computational Linguistics.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shangwen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. *arXiv preprint arXiv:2103.12953*.

Rui Zhang, Yangfeng Ji, Yue Zhang, and Rebecca J. Passonneau. 2022b. Contrastive data and learning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts*, pages 39–47, Seattle, United States. Association for Computational Linguistics.

Kun Zhou, Yuanhang Zhou, Xin Zhao, and Ji-Rong Wen. 2023. Learning to perturb for contrastive learning of unsupervised sentence representations.

A Effects of Filtering Threshold

In order to investigate the effects of the filtering threshold on the performance of our method, we analyze the results using different thresholds in Table 5. The default filtering probability for these models is set to 0.5. From the table, it can be observed that this hyperparameter has a negligible effect on the model performance. However, when the filtering threshold is set to the 90th percentile, indicating that 90% of sentences will be randomly filtered, a significant decrease in performance is observed. This outcome is expected as a high percentile threshold results in a substantial reduction in the amount of available unlabeled data, which affects the model's learning capabilities.

	Macro			Micro
	Precision	Recall	F1	F1
10-Percentile	67.6	67.6	67.5	70.5
30-Percentile	67.6	68.2	67.8	70.6
50-Percentile	**67.9**	**68.8**	**68.3**	**71.0**
70-Percentile	**67.9**	68.2	68.0	70.7
90-Percentile	66.0	67.9	66.8	70.1

Table 5: Ablation experiments with respect to the filtering threshold.

B Effects of Amount of Extra Data

	Macro			Micro
	Precision	Recall	F1	F1
0	66.2	62.3	63.4	68.7
500	64.2	63.4	63.6	67.9
1,000	66.3	64.7	65.3	69.9
2,000	68.0	64.7	65.5	69.8
3,000	67.5	65.2	66.0	70.0
5,000	67.8	66.7	66.9	70.2
8,000	67.2	68.6	67.8	70.7
10,000	**67.9**	**68.8**	**68.3**	**71.0**

Table 6: Ablation experiments with respect to the amount of unlabeled data.

In this section, we investigate the impact of the amount of unlabeled data on the performance of our method. The datasets used in these experiments are randomly sampled from a total of 10,337 unlabeled news articles that we have collected, ensuring that they have the same data distribution. *10,000* mentioned in Table 6 represents the usage of 10,337 news articles. Compared to our baseline approach where no unlabeled data is utilized, the model trained with an additional 500 unlabeled news articles does not exhibit improved performance. Since there are only 500 human-annotated articles available for training, the inclusion of 500 unlabeled articles with potential noise has a detrimental effect on the model. However, as the amount of unlabeled data increases, the model's performance gradually improves. The performances of models using 8,000 and 10,000 articles are similar, suggesting that the performance is approaching saturation.