PeerDA: Data Augmentation via Modeling Peer Relation for Span Identification Tasks

Anonymous ACL submission

Abstract

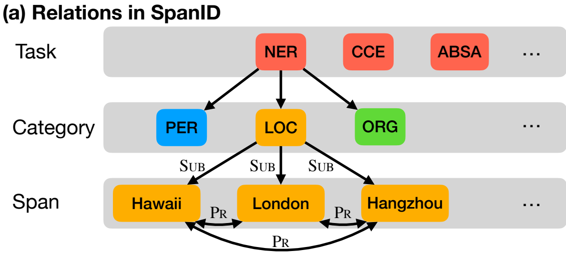
Span identification aims at identifying specific text spans from a text input and classifying them into pre-defined categories. Different from previous works that merely leverage the Subordinate (SUB) relation (i.e. *if a span is an instance of a certain category*) to train models, this paper for the first time explores the Peer (PR) relation, which indicates that *two spans are instances of the same category and share similar features*. Specifically, a novel **Peer D**ata **A**ugmentation (PeerDA) approach is proposed which employs span pairs with the PR relation as the augmentation data for training. PeerDA has two unique advantages: (1) There are a large number of PR span pairs for augmenting the training data. (2) The augmented data can prevent the trained model from overfitting the superficial span-category mapping by pushing the model to leverage the span semantics. Experimental results on ten datasets over four diverse tasks across seven domains demonstrate the effectiveness of PeerDA. Notably, PeerDA achieves state-of-the-art results on six of them.[1]

1 Introduction

Span Identification (SpanID) is a family of Natural Language Processing (NLP) tasks with the goal of detecting specific text spans and further classifying them into pre-defined categories (Papay et al., 2020). It serves as the initial step for complex text analysis by narrowing down the search scopes of important spans, which holds a pivotal position in the field of NLP (Ding et al., 2021). Recently, different domain-specific SpanID tasks, such as social media Named Entity Recognition (NER) (Derczynski et al., 2017), Aspect Based Sentiment Analysis (ABSA) (Liu, 2012), Contract Clause Extraction (CCE) (Chalkidis et al., 2017) and Span Based Propaganda Detection (SBPD) (Da San Martino et al., 2019), have emerged for various NLP applications.

[1] Our code and data are available at `github.com/XXX`.

(a) Relations in SpanID



(b) SpanID in MRC Paradigm

Context:	Gotta dress up for London fashion week and party in style!	
Original data	SUB Query:	Highlight the parts (if any) related to "LOC". Details: the name of politically or geographically defined locations such as cities, provinces, etc.
	Answer:	London
Augmented data	PR Query-1:	Highlight the parts (if any) similar to "Hawaii".
	Answer:	London
	PR Query-2:	Highlight the parts (if any) similar to "Hangzhou".
	Answer:	London

Figure 1: (a) Illustrations of Subordinate (SUB) and Peer (PR) relations in SpanID tasks. (b) The constructions of augmented data with PR relations in MRC paradigm. We use NER here for demonstration purpose.

Precisely, as shown in Figure 1 (a), the process of SpanID can be summarized as accurately extracting **span-category** Subordinate (SUB) relation — *if a span is an instance of a certain category*. Early works (Chiu and Nichols, 2016) typically tackle SpanID tasks as a sequence tagging problem, where the SUB relation is recognized via predicting the category for each input token under certain context. Recently, to better utilize category semantics, many efforts have been made on reformulating SpanID tasks as a Machine Reading Comprehension (MRC) problem (Liu et al., 2020; Yang et al., 2021). As shown by the example in Figure 1 (b), such formulation first creates a SUB query for each category and then recognizes the SUB relation by detecting relevant spans in the input text (*i.e.*, context) as answers to the category query.

However, only leveraging the SUB relation in the training data to build SpanID models may suffer from two limitations: 1) **Over-fitting**: With only SUB relation, SpanID models tend to capture the superficial span-category correlations. Such correlations may misguide the models to ignore the semantics of the given span but make predictions based on the memorized span-category patterns, which hurts the generalization capability of the models. 2) **Data Scarcity**: For low-resource scenarios or long-tailed categories, the number of span-category pairs with SUB relation (SUB pairs) could be very limited and insufficient to learn a reliable SpanID model.

In this paper, we explore the **span-span** Peer (PR) relation to alleviate the above limitations. Specifically, the PR relation indicates that *two spans are two different instances of the same category*. The major difference between PR relation and SUB relation is that the former one intends to correlate two spans without giving the categories they belong to. For example, in Figure 1 (a), "Hawaii" and "London" are connected with the PR relation because they are instances of the same category. By jointly recognizing SUB relation and PR relation in the input text, the model is enforced to favor the usage of span semantics instead of span-category patterns for prediction, reducing the risk of over-fitting. In addition, the number of span-span pairs with the PR relation (PR pairs) grows quadratically over the number of SUB pairs. Therefore, we can still construct a reasonable number of training data with PR pairs for categories having insufficient examples.

In this paper, with the aim of leveraging the PR relation to enhance SpanID models, we propose a Peer Data Augmentation (*PeerDA*) approach that treats PR pairs as a kind of augmented training data. To achieve this, as depicted in Figure 1 (b), we extend the usage of the original training data into two views. The first view is the SUB-based training data. It is used to directly solve the SpanID tasks by extracting the SUB relation, which is the typical formulation of MRC-based approaches. The second view is the PR-based training data. It is our augmentation to enrich the semantics of spans by extracting the PR relation in the original training data, where one span is used to identify its peer from the input context. Note that our PR-based training data can be easily formulated into the MRC paradigm. Therefore, the knowledge learned from such augmentation data can be directly transferred to enhance the model's capability to capture SUB relation (*i.e.*, the SpanID tasks).

To better accommodate the MRC-style SUB and PR data, we develop a stronger and more memory-efficient MRC model. Compared to the designs in Li et al. (2020b), our model introduces a bilinear component to calculate the span scores and consistently achieves better performance with a 4 times smaller memory consumption. Besides, we propose a margin-based contrastive learning strategy to additionally model the negative spans to the query (*e.g.*, when querying the context in Figure 1 for "ORG" entities, "London" becomes a negative span) so that the spans from different categories are separated more apart in the semantic space.

We evaluate the effectiveness of PeerDA on ten datasets across seven domains, from four different SpanID tasks, namely, NER, ABSA, CCE, and SBPD. Experimental results show that extracting PR relation benefits the learning of semantics and encourages models to identify more possible spans. As a result, PeerDA is a new state-of-the-art (SOTA) method on six SpanID datasets. Our analyses further demonstrate the capability of PeerDA to alleviate scarcity and over-fitting issues.

Our contributions are summarized as follows: (1) We propose a novel PeerDA approach to tackle SpanID tasks via augmenting training data with PR relation. (2) We conduct extensive experiments on ten datasets, including four different SpanID tasks across seven domains, and achieve SOTA performance on six SpanID datasets. (3) PeerDA is more effective in low-resource scenarios or long-tailed categories and thus, it alleviates the scarcity issue. Meanwhile, PeerDA pushes models to weigh more on the span semantics to prevent over-fitting.

2 Related Work

DA for SpanID: DA, which increases the diversity of training data at a low cost, is a widely-adopted solution to address data scarcity (Feng et al., 2021). In the scope of SpanID, existing DA approaches aim to introduce more span-category patterns, including: (1) *Word Replacement* that keeps the labels unchanged but replaces or paraphrases some context tokens either using simple rules (Wei and Zou, 2019; Dai and Adel, 2020) or strong language models (Kobayashi, 2018; Wu et al., 2019; Li et al., 2020a). (2) *Self-training* is to continually train the model on its predicted data (Xie et al., 2019, 2020), which shows promising results on NER (Wang et al., 2020), and propaganda detection (Hou et al., 2021). (3) *Distantly Supervised Training* focuses on leveraging external knowledge to roughly label spans in the target tasks. For example, Huang et al. (2021) leverage Wikipedia to create distant labels for NER. Chen et al. (2021) transfer data from high-resource to low-resource domains. Jain et al. (2019); Li et al. (2020c) tackle cross-lingual NER by projecting labels from high-resource to low-resource languages. Differently, the motivation of PeerDA is to leverage the augmented data to enhance models' capability on semantic understanding by minimizing(maximizing) the distances between semantically similar(distant) spans.

MRC: MRC is to extract an answer span from a relevant context conditioned on a given query. It is initially designed to solve question answering tasks (Hermann et al., 2015), while recent trends have shown great advantages of formulating NLP tasks as MRC problems. In the context of SpanID, Li et al. (2020b) address the nested NER issues by decomposing nested entities under multiple queries. Mao et al. (2021) tackle ABSA by combining aspect term extraction and sentiment polarity classification in a dual MRC framework. Hendrycks et al. (2021) tackle CCE with MRC to deal with the extraction of long clauses. Moreover, other tasks such as relation extraction (Li et al., 2019a), event detection (Liu et al., 2020, 2021), and summarization (McCann et al., 2018) are also reported to benefit from the MRC paradigm.

3 PeerDA

Overview of SpanID: Given the input text $\boldsymbol{X} = \{x_1, ..., x_n\}$, SpanID is to detect all appropriate spans $\{\boldsymbol{x}_k\}_{k=1}^{K}$ and classify them with proper labels $\{y_k\}_{k=1}^{K}$, where each span $\boldsymbol{x}_k = \{x_{s_k}, x_{s_k+1}, ..., x_{e_k-1}, x_{e_k}\}$ is a subsequence of $\boldsymbol{X}$ satisfying $s_k \leq e_k$ and the label comes from a predefined category set Y (e.g. "Person" in NER).

3.1 Training Data Construction

The training data $\mathcal{D}$ consists of two parts: (1) The SUB-based training data $\mathcal{D}^{\text{SUB}}$, where the query is about a category and the MRC context is the input text. (2) The PR-based training data $\mathcal{D}^{\text{PR}}$ is constructed with PR pairs, where one span is used to create the query and the input text containing the second span serves as the MRC context.

3.1.1 SUB-based Training Data

First, we need to transform the original training examples into (query, context, answers) triples following the paradigm of MRC (Li et al., 2020b). To extract the SUB relation between categories and relevant spans, a natural language query $\boldsymbol{Q}_y^{\text{SUB}}$ is constructed to reflect the semantics of each category y. Following Hendrycks et al. (2021), we include both category mention $[\texttt{Men}]_\text{y}$ and its definition $[\texttt{Def}]_\text{y}$ from the annotation guideline (or Wikipedia if the guideline is not accessible) in the query to introduce more comprehensive semantics:

$$\begin{aligned}\boldsymbol{Q}_y^{\text{SUB}} = &\ \texttt{Highlight the parts (if any)} \\ &\ \texttt{related to } [\texttt{Men}]_\text{y}\texttt{. Details : } [\texttt{Def}]_\text{y}.\end{aligned} \tag{1}$$

Given the input text $\boldsymbol{X}$ as the context, the answers to $\boldsymbol{Q}_y^{\text{SUB}}$ are the spans belonging to category y. Then we can obtain one MRC example denoted as $(\boldsymbol{Q}_y^{\text{SUB}}, \boldsymbol{X}, \{\boldsymbol{x}_k \mid \boldsymbol{x}_k \in \boldsymbol{X}, y_k = y\}_{k=1}^{K})$. To guarantee the identification of all possible spans, we create $|Y|$ training examples by querying the input text with each pre-defined category.

3.1.2 PR-based training data

To construct augmented data that derived from the PR relation, we first create a category-wise span set $\mathcal{S}_y$ that includes all training spans with category y:

$$\mathcal{S}_y = \{\boldsymbol{x}_k \mid (\boldsymbol{x}_k, y_k) \in \mathcal{D}^{\text{SUB}}, y_k = y\} \tag{2}$$

Obviously, any two different spans in $\mathcal{S}_y$ have the same category and shall hold the PR relation. Therefore, we pair every two different spans in $\mathcal{S}_y$ to create a peer set $\mathcal{P}_y$:

$$\mathcal{P}_y = \{(\boldsymbol{x}^q, \boldsymbol{x}^a) \mid \boldsymbol{x}^q, \boldsymbol{x}^a \in \mathcal{S}_y, \boldsymbol{x}^q \neq \boldsymbol{x}^a\} \tag{3}$$

For each PR pair $(\boldsymbol{x}^q, \boldsymbol{x}^a)$ in $\mathcal{P}_y$, we can construct one training example by constructing the query with the first span $\boldsymbol{x}^q$:

$$\begin{aligned}\boldsymbol{Q}_y^{\text{PR}} = &\ \texttt{Highlight the parts (if any)} \\ &\ \texttt{similar to } \boldsymbol{x}^q.\end{aligned} \tag{4}$$

Then we treat the text $\boldsymbol{X}^a$ containing the second span $\boldsymbol{x}^a$ as the MRC context to be queried and $\boldsymbol{x}^a$ as the answer to $\boldsymbol{Q}_y^{\text{PR}}$. Note that there may exist more than one span in $\boldsymbol{X}^a$ satisfying PR relation with $\boldsymbol{x}^q$, we set all of them as the valid answers to $\boldsymbol{Q}_y^{\text{PR}}$, yielding one training example $(\boldsymbol{Q}_y^{\text{PR}}, \boldsymbol{X}^a, \{\boldsymbol{x}_k^a \mid \boldsymbol{x}_k^a \in \boldsymbol{X}^a, y_k^a = y\}_{k=1}^{K})$ of our PeerDA.

Theoretically, given the span set $\mathcal{S}_y$, there are only $|\mathcal{S}_y|$ SUB pairs in the training data but we can obtain $|\mathcal{S}_y| \times (|\mathcal{S}_y| - 1)$ PR pairs to construct $\mathcal{D}^{\text{PR}}$. Such a large number of augmented data shall hold great potential to enrich spans' semantics. However, putting all PR-based examples into training would exacerbate the skewed data distribution issue since the long-tailed categories get fewer PR pairs for augmentation and also increase the training cost. Therefore, as the first step for DA with the PR relation, we propose three augmentation strategies to control the size and distribution of augmented data.

PeerDA-Size: This is to increase the size of augmented data while keeping the data distribution unchanged. Specifically, for each category y, we randomly sample $\lambda|\mathcal{S}_y|$ PR pairs from $\mathcal{P}_y$. Then we collect all sampled PR pairs to construct $\mathcal{D}^{\text{PR}}$, where λ is the DA rate to control the size of $\mathcal{D}^{\text{PR}}$.

PeerDA-Categ: Categories are not evenly distributed in the training data, and in general SpanID models perform poorly on long-tailed categories. To tackle this, we propose PeerDA-Categ to augment more training data for long-tailed categories. Specifically, let y^* denote the category having the largest span set of size $|\mathcal{S}_{y^*}|$. We sample up to $|\mathcal{S}_{y^*}| - |\mathcal{S}_y|$ PR pairs from $\mathcal{P}_y$ for each category y and construct a category-balanced training set $\mathcal{D}^{\text{PR}}$ using all sampled pairs. Except for the extreme cases where $|\mathcal{S}_y|$ is smaller than $\sqrt{|\mathcal{S}_{y^*}|}$, we would get the same size of the training data for each category after the augmentation, which significantly increases the exposure for spans from the long-tailed categories.

PeerDA-Both (The final version of PeerDA): To take advantage of the above two strategies, we further propose PeerDA-Both to maintain the data distribution while effectively increasing the size of training data. In PeerDA-Both, we randomly sample $\max(\lambda|\mathcal{S}_{y^*}| + (|\mathcal{S}_{y^*}| - |\mathcal{S}_y|), 0)$ PR pairs from $\mathcal{P}_y$ for each category y to construct $\mathcal{D}^{\text{PR}}$, where $\lambda|\mathcal{S}_{y^*}|$ determines the size of the augmented data, and $|\mathcal{S}_{y^*}| - |\mathcal{S}_y|$ controls the data distribution.

3.1.3 Data Balance

We combine the $\mathcal{D}^{\text{SUB}}$ and the $\mathcal{D}^{\text{PR}}$ created above as the final training data. Since an input text usually mentions spans from a few categories, when converting the text into the MRC paradigm, many of the $|Y|$ examples are unanswerable. If a SpanID model is trained on this unbalanced data, then the model may favor the majority of the training examples and output an empty span. To balance answerable and unanswerable examples, we follow Hendrycks et al. (2021) to randomly remove some unanswerable examples from the training data.

3.2 Model Architecture

As shown in Figure 2, to achieve the detection of multiple spans for the given query, we follow Li et al. (2020b) to build the MRC model. Compared to the original designs, we further optimize the computation of span scores following a *general* way of Luong et al. (2015).

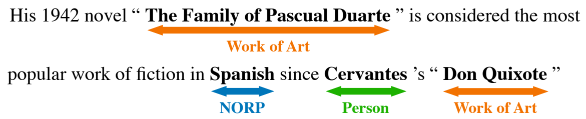


Figure 2: Example of extracting multiple spans in NER.

Specifically, the base model consists of three components: an encoder, a span predictor, and a start-end selector. First, given the concatenation of the query $\boldsymbol{Q}$ and the context $\boldsymbol{X}$ as the MRC input $\overline{\boldsymbol{X}} = \{[\texttt{CLS}], \boldsymbol{Q}, [\texttt{SEP}], \boldsymbol{X}, [\texttt{SEP}]\}$, where $[\texttt{CLS}]$, $[\texttt{SEP}]$ are special tokens, the encoder would encode the input text into hidden states $\boldsymbol{H}$:

$$\boldsymbol{H} = \textsc{Encoder}(\overline{\boldsymbol{X}}) \tag{5}$$

Second, the span predictor consists of two binary classifiers, one to predict whether each context token is the start index of the answer, and the other to predict whether the token is the end index:

$$P_{\text{start}} = \boldsymbol{H}W^s \qquad P_{\text{end}} = \boldsymbol{H}W^e \tag{6}$$

where $W^s, W^e \in \mathbb{R}^{d\times 2}$ are the weights of two classifiers and d is the dimension of hidden states. The span predictor would output multiple start and end indexes for the given query and context.

Third, the start-end selector matches each start index to each end index and selects the most possible spans from all combinations as the outputs. Different from the *concat* way that would create a large $\mathbb{R}^{|\overline{\boldsymbol{X}}|\times|\overline{\boldsymbol{X}}|\times 2d}$-shape tensor (Li et al., 2020b), we leverage a *general* way following Luong et al. (2015) to compute the span score, consuming fewer resources for better training efficiency:

$$P_{\text{s,e}} = \mathit{FFN}(\boldsymbol{H}_s)^T \boldsymbol{H}_e \tag{7}$$

where *FFN* is the feed-forward network (Vaswani et al., 2017), $P_{\text{s,e}}$ denotes the likelihood of $\overline{\boldsymbol{X}}_{s:e}$ to form a possible answer.

3.3 Training Objective

The standard objective is to minimize the cross-entropy loss (CE) between above three predictions and their corresponding ground-truth labels, i.e., $Y_{\text{start}}, Y_{\text{end}}, Y_{\text{s,e}}$ (Li et al., 2020b):

$$\begin{aligned}\mathcal{L}_{mrc} = \text{CE}(\sigma(P_{\text{start}}), Y_{\text{start}}) + \text{CE}(\sigma(P_{\text{end}}), Y_{\text{end}}) \\ + \text{CE}(\sigma(P_{\text{s,e}}), Y_{\text{s,e}})\end{aligned} \tag{8}$$

where σ is the sigmoid function.

However, these objectives only capture the semantic similarity between the query and positive

Task	NER					ABSA		SBPD		CCE
Dataset	**OntoNotes5**	**WNUT17**	**Movie**	**Restaurant**	**Weibo**	**Lap14**	**Rest14**	**News20**	**Social21**	**CUAD**
Domain	*mixed*	*social*	*movie*	*restaurant*	*social*	*laptop*	*restaurant*	*news*	*social*	*legal*
# Train	60.0k	3.4k	7.8k	7.7k	1.3k	2.7k	2.7k	0.4k	0.7k	0.5k
# Test	8.3k	1.3k	2.0k	1.5k	0.3k	0.8k	0.8k	75 (*dev*)	0.2k	0.1k
# Category	11	6	12	8	4	1 / 3	1 / 3	14	20	41

Table 1: Statistics on the ten SpanID datasets. Note that 1 / 3 denotes that there is 1 category in ATE and 3 categories in UABSA. *dev* denotes that we evaluate News20 on the dev set.

spans (i.e., the span instances of the query category). In this paper, we propose to explicitly separate the query and its negative spans (i.e., the span instances of other categories) apart with a margin-based contrastive learning strategy, for better distinguishing the spans from different categories.

Specifically, given the MRC input $\overline{\boldsymbol{X}}$ with query of category y, there may be multiple positive spans $\overline{\mathcal{X}}^+ = \{\overline{\boldsymbol{x}}_k \in \overline{\boldsymbol{X}}, y_k = y\}$ and negative spans $\overline{\mathcal{X}}^- = \{\overline{\boldsymbol{x}}_{k'} \in \overline{\boldsymbol{X}}, y_{k'} \neq y\}$. We leverage the following margin-based contrastive loss to penalize negative spans (Chechik et al., 2010):

$$\mathcal{L}_{ct} = \max_{\substack{\overline{\boldsymbol{x}}_k \in \overline{\mathcal{X}}^+ \\ \overline{\boldsymbol{x}}_{k'} \in \overline{\mathcal{X}}^-}} max(0, M - (\sigma(P_{s_k,e_k}) - \sigma(P_{s_{k'},e_{k'}}))) \tag{9}$$

where M is the margin term, $max(\cdot,\cdot)$ is to select the larger one from two candidates, and the span score P_{s_k,e_k} can be regarded as the semantic similarity between the query and the target span $\overline{\boldsymbol{x}}_k$. Note that our contrastive loss maximizes the similarity difference between the query and the most confusing positive and negative span pairs (*Max-Min*), which we demonstrate to be effective in Sec. 5.3.

Finally, the overall training objective is:

$$\mathcal{L} = \mathcal{L}_{mrc} + \alpha \mathcal{L}_{ct} \tag{10}$$

where α is the balance rate.

4 Experimental Setup

4.1 Tasks

We conduct experiments on four SpanID tasks from diverse domains, including NER, ABSA, Contract Clause Extraction (CCE), and Span Based Propaganda Detection (SBPD). The dataset statistics are summarized in Table 1. The detailed task description can be found in Appendix A.1.

NER: We evaluate five datasets, including four English datasets: **OntoNotes5**[2] (Pradhan et al., 2013), **WNUT17** (Derczynski et al., 2017), **Movie** (Liu et al., 2013b), and **Restaurant** (Liu et al., 2013a) and a Chinese dataset **Weibo** (Peng and Dredze, 2015). We use micro-averaged Precision, Recall, and F_1 as evaluation metrics.

ABSA: We explore two ABSA sub-tasks: **Aspect Term Extraction (ATE)** to only extract aspect terms, and **Unified Aspect Based Sentiment Analysis (UABSA)** to jointly identify aspect terms and their sentiment polarities. We evaluate the two sub-tasks on two datasets, including the laptop domain **Lap14** and restaurant domain **Rest14**. We use micro-averaged F_1 as the evaluation metric.

SBPD: It aims to detect both the text fragment where a persuasion technique is used and its technique type. We use **News20** and **Social21** from SemEval shared tasks (Da San Martino et al., 2020; Dimitrov et al., 2021). For **News20**, we report the results on its dev set since the test set is not publicly available. We use micro-averaged Precision, Recall, and F_1 as evaluation metrics.

CCE: It is a legal task to detect and classify contract clauses into relevant clause types, such as "Governing Law". We conduct CCE experiments using **CUAD** (Hendrycks et al., 2021). We follow Hendrycks et al. (2021) to use Area Under the Precision-Recall Curve (AUPR) and Precision at 80% Recall (P@0.8R) as the evaluation metrics.

4.2 Implementations

Since legal SpanID tasks have a lower tolerance for missing important spans, we do not include start-end selector (i.e. CE($P_{\text{s,e}}, Y_{\text{s,e}}$) and $\alpha\mathcal{L}_{ct}$ in Eq. (10)) in the CCE models but follow Hendrycks et al. (2021) to output top 20 spans from span predictor for each input example in order to extract spans as much as possible. While for NER, ABSA, and SBPD, we use our optimized architecture and objective. For fair comparison with existing works, our models utilize BERT (Devlin et al., 2019) as the text encoder for ABSA and RoBERTa (Liu et al., 2019) for NER, CCE, and SBPD. Detailed

[2] In order to conduct robustness experiments in Sec. A.4, we use the datasets from Lin et al. (2021) with 11 entity types.

Methods	OntoNotes5			WNUT17			Movie			Restaurant			Weibo		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
	RB-CRF+RM			CL-KL			T-NER			KaNa			RoBERTa+BS		
SOTA	92.8	92.4	92.6	-	-	**60.5**	-	-	71.2	80.9	80.0	80.4	70.2	75.4	**72.7**
							Base								
Tagging	91.0	91.8	91.4	62.1	48.2	54.3	73.0	72.8	72.9	80.6	80.7	80.7	70.8	71.0	70.9
MRC	92.4	91.8	92.1	66.4	40.7	50.5	70.3	73.3	71.8	81.4	79.9	80.6	73.6	64.4	68.7
PeerDA	91.9	92.6	92.4	71.1	46.9	56.5	77.9	72.3	75.0	81.3	82.8	82.1	70.0	73.3	71.6
							Large								
Tagging	93.0	92.3	92.6	69.4	46.2	55.4	74.2	74.0	74.1	80.9	82.0	81.4	71.4	69.2	70.3
MRC	92.8	91.8	92.3	72.4	41.7	52.9	76.7	73.2	74.9	81.6	81.7	81.7	72.2	66.8	69.4
PeerDA	92.8	93.7	**93.3**	70.9	48.0	57.2	78.5	73.1	**75.7**	81.8	82.5	**82.2**	73.4	71.6	72.5

Table 2: Performance on NER datasets. The best models are bolded.

configurations can be found in Appendix A.

4.3 Baselines

Note that our main contribution is to provide a new perspective to treat the PR relation as a kind of training data for augmentation. Therefore, we compare with models built on the same encoder-only PLMs (Devlin et al., 2019; Liu et al., 2019). We are not focusing on pushing the SOTA results to new heights though some of the baselines already achieved SOTA performance.

NER: We compare with Tagging (Liu et al., 2019) and MRC (Li et al., 2020b) baselines. We also report the previous best approaches for each dataset, including RB-CRF+RM (Lin et al., 2021), CL-KL (Wang et al., 2021), T-NER (Ushio and Camacho-Collados, 2021) KaNa (Nie et al., 2021), and RoBERTa+BS (Zhu and Li, 2022).

ABSA: In addition to MRC baseline, we also compare with previous approaches on top of BERT. These are SPAN-BERT (Hu et al., 2019), IMN-BERT (He et al., 2019), RACL (Chen and Qian, 2020) and Dual-MRC (Mao et al., 2021).

SBPD: For **News20** we only compare with MRC baseline due to the lack of related work. For **Social21**, we compare with top three approaches on its leaderboard, namely, Volta (Gupta et al., 2021), HOMADOS (Kaczyński and Przybyła, 2021), and TeamFPAI (Hou et al., 2021).

CCE: We compare with (1) MRC basline, (2) stronger text encoders, including ALBERT (Lan et al., 2019) and DeBERTa (He et al., 2020), and (3) the model continually pretrained on contracts: RoBERTa + CP (Hendrycks et al., 2021).

Methods	Lap14		Rest14	
	UABSA	ATE	UABSA	ATE
SPAN-BERT	61.3	82.3	73.7	86.7
IMN-BERT	61.7	77.6	70.7	84.1
RACL	63.4	81.8	75.4	86.4
Dual-MRC	65.9	82.5	**76.0**	86.6
MRC (Large)	63.2	83.9	72.9	86.8
PeerDA	**65.9**	**84.6**	73.9	**86.8**

Table 3: Performance on two ABSA subtasks on two datasets. Results are averages F_1 over 5 runs.

5 Results

5.1 Comparison Results

NER: Table 2 shows the performance on five NER datasets. Our PeerDA significantly outperforms the Tagging and MRC baselines. Precisely, compared to RoBERTa$_{\texttt{base}}$ MRC, PeerDA obtains 0.3, 6.0, 3.2, 1.5, and 2.9 F_1 gains on five datasets respectively. When implemented on RoBERTa$_{\texttt{large}}$, our PeerDA can further boost the performance and establishes new SOTA on three datasets, namely, **OntoNotes5**, **Movie**, and **Restaurant**. Note that the major improvement of PeerDA over MRC comes from higher Recall. It implies that PeerDA encourages models to give more span predictions.

ABSA: Table 3 depicts the results on ABSA. Compared to previous approaches, PeerDA mostly achieves better results on two subtasks, where it outperforms vanilla MRC by 2.7 and 1.0 F_1 on UABSA for two domains respectively.

SBPD: The results of two SBPD tasks are presented in Table 4. PeerDA outperforms MRC by 8.2 and 9.2 F_1 and achieves SOTA performance on **News20** and **Social21** respectively.

CCE: The results of CCE are shown in Table 5. PeerDA surpasses MRC by 8.7 AUPR and 13.3

Methods	News20			Social21		
	P	R	F_1	P	R	F_1
Volta	-	-	-	50.1	46.4	48.2
HOMADOS	-	-	-	41.2	40.3	40.7
TeamFPAI	-	-	-	65.2	28.6	39.7
MRC (Base)	10.5	53.5	17.6	55.8	43.5	48.9
PeerDA	21.8	31.5	**25.8**	49.4	70.6	**58.1**

Table 4: PeerDA performance on two SBPD datasets.

Methods	#Params	AUPR	P@0.8R
$\text{ALBERT}_{\texttt{xxlarge}}$	223M	38.4	31.0
$\text{RoBERTa}_{\texttt{base}}$ + CP	125M	45.2	34.1
$\text{RoBERTa}_{\texttt{large}}$	355M	48.2	38.1
$\text{DeBERTa}_{\texttt{xlarge}}$	900M	47.8	44.0
MRC (Base)	125M	43.6	32.2
PeerDA	125M	**52.3**	45.5

Table 5: PeerDA performance on CCE.

Ablation Type	NER	UABSA	SBPD	CCE	Avg.
MRC	72.7	68.1	33.3	43.6	54.4
PeerDA-Size	74.6	69.7	38.5	48.7	57.9
PeerDA-Categ	74.2	69.3	40.4	51.3	58.8
PeerDA-Both (**final**)	**75.5**	**69.9**	**42.0**	**52.3**	**59.9**

Table 6: Ablation study on data augmentation strategies. The results (F_1 for NER, UABSA, and SBPD. AUPR for CCE) are averaged of all datasets in each task.

Ablation Type	|GPU|	NER	UABSA	SBPD	Avg.
Calculation of $P_{s,e}$					
concat	1x	74.5	69.2	40.3	61.3
general (**final**)	**0.23x**	**75.0**	**69.4**	**40.8**	**61.7**
Contrastive Loss					
Average	0.23x	75.1	69.6	37.6	60.8
Max-Min (**final**)	0.23x	**75.5**	**69.9**	**42.0**	**62.4**

Table 7: Ablation study on model designs. The F_1 scores are averaged of all datasets in each task. The **|GPU|** column denotes the GPU memory footprint of each variant under the same experimental setup.

P@0.8R and even surpasses the model of extremely large size ($\text{DeBERTa}_{\texttt{xlarge}}$) by 4.5 AUPR, reaching SOTA performance on **CUAD**.

5.2 Analysis on Augmentation Strategies

To explore how the size and category distribution of the augmented data affect the SpanID tasks, we conduct ablation study on the three augmentation strategies mentioned in Sec. 3.1.2, depicted in Table 6. Overall, all of the PeerDA variants are clearly superior to the MRC baseline and the PeerDA-both considering both data size and distribution issues performs the best. Another interesting finding is that PeerDA-Categ significantly outperforms PeerDA-Size on SBPD and CCE. We attribute the phenomenon to the fact that SBPD and CCE have a larger number of categories and consequently, the MRC model is more prone to the issue of skewed data distribution. Under this circumstance, PeerDA-Categ, the variant designed for compensating the long-tailed categories, can bring larger performance gains over MRC model. On the other hand, if the skewed data distribution is not severe (e.g. NER), or the category shows a weak correlation with the spans (i.e. UABSA), PeerDA-Size is more appropriate than PeerDA-Categ.

5.3 Analysis on Model Designs

Calculation of $P_{s,e}$ (Top part of Table 7) Under the same experimental setup ($\text{RoBERTa}_{\texttt{base}}$, batch size=32, sequence length=192, fp16), using our *general* method (Eq. (7)) to compute span score $P_{s,e}$ greatly reduces the memory footprint by more than 4 times with no performance drop, compared to the original *concat* method. Therefore, our *general* method allows a larger batch size for accelerating the training.

Contrastive Loss (Bottom part of Table 7) After we have settled on the *general* scoring function, we further investigate different methods to compute contrastive loss. We find that the *Average* method, which averages similarity differences between the query and all pairs of positive and negative spans, would affect SpanID performance when the task has more long-tailed categories (i.e. SBPD). While our *Max-Min* (strategy in Eq.(9)) is a relaxed regularization, which empirically is more suitable for SpanID tasks and consistently performs better than the *Average* method.

6 Further Discussions

In this section, we make further discussions to bring valuable insights of our PeerDA approach.

Out-of-domain Evaluation: We conduct out-of-domain evaluation on four English NER datasets, where the model is trained on **OntoNotes5**, the largest dataset among them, and evaluated on the test part of another three datasets. Since these four datasets are from different domains and differ substantially in their categories, this setting largely eliminates the impact of superficial span-category patterns and thus it can faithfully reflect how well

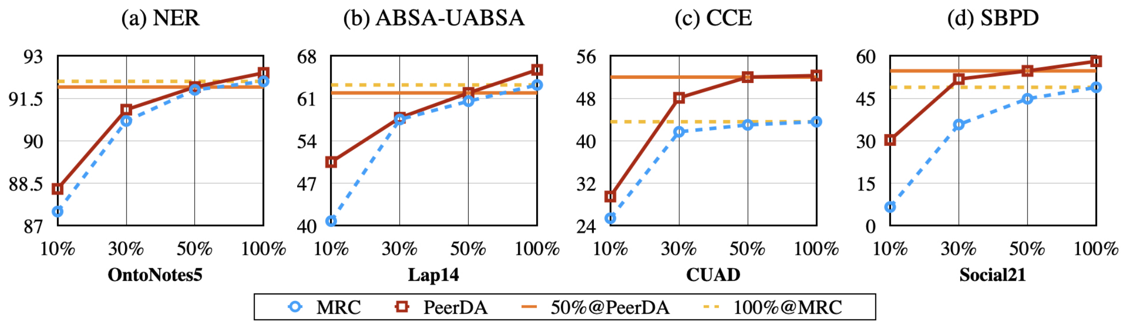


Figure 3: Performance on low-resource scenarios. We select one dataset for each SpanID task and report the test results (AUPR for CCE and F_1 for others) from the models trained on different proportions of the training data.

SRC → TGT	RoBERTa$_{\text{base}}$		RoBERTa$_{\text{large}}$	
	MRC	PeerDA	MRC	PeerDA
Onto. → WNUT17	43.1	**46.8**	44.2	**46.9**
Onto. → Rest.	1.6	**5.0**	2.7	**11.0**
Onto. → Movie	25.0	**26.7**	26.7	**27.8**
Average	23.3	**26.2**	24.5	**28.6**

Table 8: F_1 scores on NER cross-domain transfer, where models trained on source-domain training data (SRC) are evaluated on target-domain test sets (TGT).

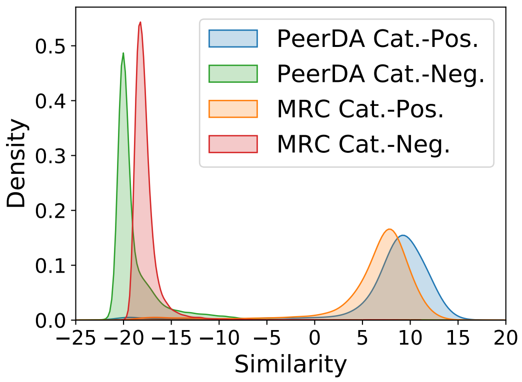


Figure 4: The distribution of similarity score between categories and their corresponding positive/negative spans on **Ontonotes5** test set.

the MRC model exploits span semantics for prediction. The results are presented in Table 8. PeerDA can significantly exceed MRC on all three transfer pairs. On average, PeerDA achieves 2.9 and 4.1 F_1 gains over base-size MRC and large-size MRC respectively. These results verify our postulation that modeling the PR relation allows models to weigh more on the semantics for making predictions, and thus mitigates the over-fitting issue.

Semantic Distance: To gain a deeper understanding of the way in which PeerDA enhances model performance, we consider the span score (Eq. 7) as a measure of semantic similarity between a query and a span. In this context, we can create queries for all categories and visualize the similarity distribution between the categories and their corresponding positive and negative spans on **Ontonote5** test set. As shown in Figure 4, we can observe that the use of PeerDA leads to an increased semantic similarity between spans and their corresponding categories, resulting in higher confidence in the prediction of correct spans. Furthermore, PeerDA has been shown to also create a larger similarity gap between positive and negative spans, facilitating their distinction.

Low-resource Evaluation: We simulate low-resource scenarios by randomly selecting 10%, 30%, 50%, and 100% of the training data for training SpanID models and show the comparison results between PeerDA and MRC on four SpanID tasks in Figure 3. As can be seen, our PeerDA further enhances the MRC model in all sizes of training data and the overall trends are consistent across the above four tasks. When training PeerDA with 50% of the training data, it can reach or even exceed the performance of MRC trained on the full training set. These results demonstrate the effectiveness of our PeerDA in low-resource scenarios.

7 Conclusions

In this paper, we propose a novel PeerDA approach for SpanID tasks to augment training data from the perspective of capturing the PR relation. PeerDA has two unique advantages: (1) It is capable to leverage abundant but previously unused PR relation as additional training data. (2) It alleviates the over-fitting issue of MRC models by pushing the models to weigh more on semantics. We conduct extensive experiments to verify the effectiveness of PeerDA. Further in-depth analyses demonstrate that the improvement of PeerDA comes from a better semantic understanding capability.

Limitations

In this section, we discuss the limitations of this work as follows:

- PeerDA leverages labeled spans in the existing training set to conduct data augmentation. This means that PeerDA improves the semantics learning of existing labeled spans, but is ineffective to classify other spans outside the training set. Therefore, it would be beneficial to engage outer source knowledge (e.g. Wikipedia), where a variety of important entities and text spans can also be better learned with our PeerDA approach.
- Since PeerDA is designed in the MRC formulation on top of the encoder-only Pre-trained Language Models (PLMs) (Devlin et al., 2019; Liu et al., 2019), it is not comparable with other methods built on encoder-decoder PLMs (Yan et al., 2021b; Chen et al., 2022; Zhang et al., 2021; Yan et al., 2021a). It would be of great value to try PeerDA on encoder-decoder PLMs such as BART (Lewis et al., 2020) and T5 (Raffel et al., 2020), to see whether PeerDA is a general approach regardless of model architecture.
- As shown in Table 12, although PeerDA can significantly alleviate the *Missing Predictions*, the most prevailing error in the MRC model, PeerDA also introduces some new errors, i.e. *Multiple labels* and *Incorrect Label*. It should be noted that those problematic spans are usually observed in different span sets, where they would learn different category semantics from their peers. Therefore, we speculate that those spans tend to leverage the learned category semantics more than their context information to determine their categories. We hope such finding can shed light on future research to further improve PeerDA.

References

Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. 2017. Extracting contract elements. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 19–28.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research*, 11(3).

Shuguang Chen, Gustavo Aguilar, Leonardo Neves, and Thamar Solorio. 2021. Data augmentation for cross-domain named entity recognition. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5346–5356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Chen, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, Huajun Chen, and Ningyu Zhang. 2022. LightNER: A lightweight tuning paradigm for low-resource NER via pluggable prompting. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2374–2387, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694, Online. Association for Computational Linguistics.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 140–147, Copenhagen, Denmark. 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.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. Volta at SemEval-2021 task 6: Towards detecting persuasive texts and images using textual and multimodal ensemble. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1075–1081, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pages 1693–1701.

Xiaolong Hou, Junsong Ren, Gang Rao, Lianxin Lian, Zhihao Ruan, Yang Mo, and JIanping Shen. 2021. FPAI at SemEval-2021 task 6: BERT-MRC for propaganda techniques detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1056–1060, Online. Association for Computational Linguistics.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Florence, Italy. Association for Computational Linguistics.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-shot named entity recognition: An empirical baseline study. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10408–10423, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alankar Jain, Bhargavi Paranjape, and Zachary C. Lipton. 2019. Entity projection via machine translation for cross-lingual NER. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1083–1092, Hong Kong, China. Association for Computational Linguistics.

Konrad Kaczyński and Piotr Przybyła. 2021. HOMADOS at SemEval-2021 task 6: Multi-task learning for propaganda detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1027–1031, Online. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020a. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066, Online. Association for Computational Linguistics.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020b. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859, Online. Association for Computational Linguistics.

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019a. Entity-relation extraction as multi-turn question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1340–1350, Florence, Italy. Association for Computational Linguistics.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019b. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6714–6721.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020c. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. RockNER: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Jian Liu, Yufeng Chen, and Jinan Xu. 2021. Machine reading comprehension as data augmentation: A case study on implicit event argument extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2725, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingjing Liu, Panupong Pasupat, Scott Cyphers, and Jim Glass. 2013a. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8386–8390. IEEE.

Jingjing Liu, Panupong Pasupat, Yining Wang, Scott Cyphers, and Jim Glass. 2013b. Query understanding enhanced by hierarchical parsing structures. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 72–77. IEEE.

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 preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *arXiv preprint arXiv:2101.00816*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Binling Nie, Ruixue Ding, Pengjun Xie, Fei Huang, Chen Qian, and Luo Si. 2021. Knowledge-aware named entity recognition with alleviating heterogeneity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13595–13603.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1383–1391, Online. Association for Computational Linguistics.

Sean Papay, Roman Klinger, and Sebastian Padó. 2020. Dissecting span identification tasks with performance prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4881–4895, Online. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

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.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Improving named entity recognition by external context retrieving and cooperative learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1800–1812, Online. Association for Computational Linguistics.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2020. Adaptive self-training for few-shot neural sequence labeling. *arXiv preprint arXiv:2010.03680*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuanjing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605, Online. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021a. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021b. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5808–5822, Online. Association for Computational Linguistics.

Pan Yang, Xin Cong, Zhenyu Sun, and Xingwu Liu. 2021. Enhanced language representation with label knowledge for span extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4623–4635, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Task Overview

We conduct experiments on four SpanID tasks with diverse domains, including Named Entity Recognition (NER), Aspect Based Sentiment Analysis (ABSA), Contract Clause Extraction (CCE) and Span Based Propaganda Detection (SBPD), to show the overall effectiveness of our PeerDA. The dataset statistics are summarized in Table 1.

NER is a traditional SpanID task, where spans denote the named entities in the input text and category labels denote their associated entity types. We evaluate five datasets from four domains:

- **OntoNotes5** (Pradhan et al., 2013) is a large-scale mixed domain NER dataset covering News, Blog and Dialogue. To make a fair comparison in the robustness experiments in Sec. A.4, we use the datasets from Lin et al. (2021), which only add adversarial attack to the 11 entity types, while leaving out 7 numerical types.

- **WNUT17** (Derczynski et al., 2017) is a benchmark NER dataset in social media domain. For fair comparison, we follow the data preprocessing protocols in Nie et al. (2020).

- **Movie** (Liu et al., 2013b) is a movie domain dataset containing movie queries, where long spans are annotated such as a movie's origin or plot. We use the defaulted data split strategy into train, test sets.

- **Restaurant** (Liu et al., 2013a) contains queries in restaurant domain. Similar to Movie, we use the defaulted data split strategy.

- **Weibo** (Peng and Dredze, 2015) is a Chinese benchmark NER dataset in social media domain. We exactly follow the official data split strategy into train, dev and test sets.

ABSA (Li et al., 2019b; Chen and Qian, 2020) is a fine-grained sentiment analysis task centering at aspect terms. We explore two ABSA sub-tasks:

- **Aspect Term Extraction (ATE)** is to extract aspect terms, where there is only one query asking if there are any aspect terms in the input text.

- **Unified Aspect Based Sentiment Analysis (UABSA)** is to jointly extract aspect terms and predict their sentiment polarities. We formulate it as a SpanID task by treating the sentiment polarities, namely, positive, negative, and neutral, as three category labels, and aspect terms as spans.

We evaluate the two sub-tasks on two datasets, including the laptop domain dataset **Lap14** and restaurant domain dataset **Rest14** from SemEval Shared tasks (Pontiki et al., 2014). We use the processed data from Zhang et al. (2021).

CCE is a legal NLP task to detect and classify contract clauses into relevant clause types, such as "Governing Law" and "Uncapped Liability". The goal of CCE is to reduce the labor of legal professionals in reviewing contracts of dozens or hundreds of pages long. CCE is also a kind of SpanID task where spans are those contract clauses that warrant review or analysis and labels are predefined clause types. We conduct experiments on CCE using **CUAD** (Hendrycks et al., 2021), where they annotate contracts from Electronic Data Gathering, Analysis and Retrieval (EDGAR) with 41 clause types. We follow Hendrycks et al. (2021) to split the contracts into segments within the length limitation of pretrained language models and treat each individual segment as one example. We also follow their data split strategy.

SBPD (Da San Martino et al., 2019) is a typical SpanID task that aims to detect both the text fragment (i.e. spans) where a persuasion technique is being used as well as its technique type (i.e. category labels). We use the **News20** and **Social21** from two SemEval shared tasks (Da San Martino et al., 2020; Dimitrov et al., 2021) and follow the official data split strategy. Note that **News20** does not provide the golden label for the test set. Therefore, we evaluate **News20** on the dev set.

A.2 Implementations

We use Huggingface's implementations of BERT and RoBERTa (Wolf et al., 2020) [3]. The hyperparameters can be found in Table 9. We use Tesla V100 GPU cards for conducting all the experiments. We follow the default learning rate schedule and dropout settings used in BERT. We use AdamW (Loshchilov and Hutter, 2019) as our optimizer. The margin term M is set to 0 for NER and ABSA, and 1 for SBPD. The balance rate α is set to 0.1.

[3]Chinese RoBERTa is from https://github.com/ymcui/Chinese-BERT-wwm.

Dataset	OnteNote5	WNUT17	Movie	Restaurant	Weibo	Lap14	Rest14	CUAD	News20	Social21
Query Length	32	32	64	64	64	24	24	256	80	80
Input Length	160	160	160	128	192	128	128	512	200	200
Batch Size	32	32	32	32	8	16	16	16	16	16
Learning Rate	2e-5	1e-5	1e-5	1e-5	1e-5	2e-5	2e-5	5e-5	2e-5	3e-5
λ	1	1	1	1	1	1	1	-0.5	0.5	1

Table 9: Hyper-parameters settings.

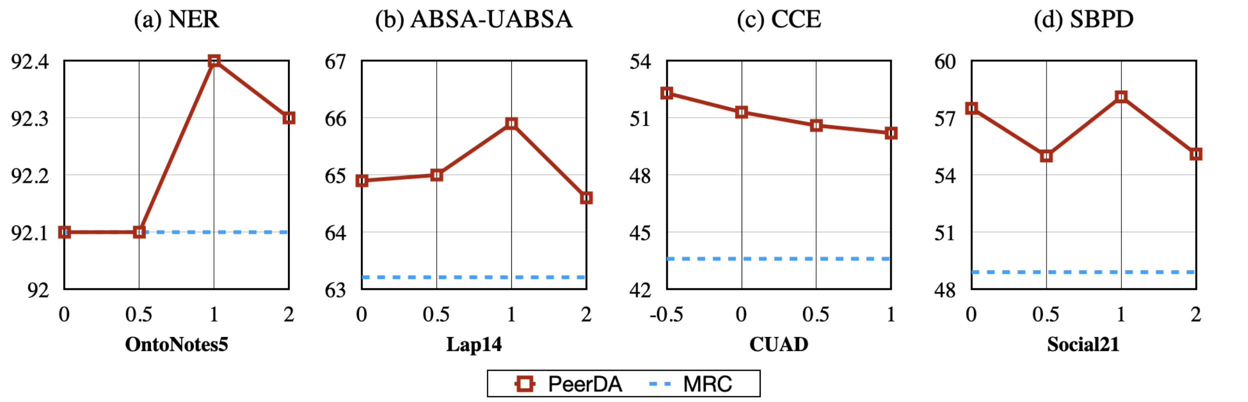


Figure 5: Performance in terms of different DA rate λ. We vary λ to get different volumes of PR-based training data.

A.3 Effect of DA Rate

We vary the DA rate λ to investigate how the volume of PR-based training data affect the SpanID models performance.

Figure 5 shows the effect of different λ in four SpanID tasks. PeerDA mostly improves the MRC in all different trials of λ and we suggest that some parameter tuning for λ is beneficial to obtain optimal results.

Another observation is that too large λ would do harm to the performance. Especially on CCE, due to the skewed distribution and a large number of categories, PeerDA can produce a huge size of PR-based training data. We speculate that too much PR-based training data would affect the learning of BL-based training data and thus affect the model's ability to solve a SpanID task, causing the optimal λ to be a negative value. In addition, too much PR-based training data would also increase the training cost. As a result, we should maintain an appropriate ratio of BL-based and PR-based training data to keep a reasonable performance on SpanID tasks.

A.4 Robustness:

To verify the advantage of PeerDA against the adversarial attack, we conduct robustness experiments using the adversarial dev set of **OntoNotes5** (Lin et al., 2021) on NER and adversarial test set of **Lap14** (Xing et al., 2020) on UABSA. Table 10 shows the performance on the original and the adversarial sets. On **OntoNotes5** *full* adversarial set, PeerDA improves the robustness of the model compared to MRC but slightly degrades compared to Tagging. To investigate why this happens, we evaluate each type of adversarial attack independently, including *entity* attack that replaces entities to other entities not presented in the training set and *context* attack that replaces the context of entities. It shows that PeerDA does not work well on *entity* attack because we only use entities in the training set to conduct data augmentation, which is intrinsically ineffective to this adversarial attack. This motivates us to engage outer source knowledge (e.g. Wikipedia) into our PeerDA approach in future work. On **Lap14**, PeerDA significantly improves Tagging and MRC by 5.6 and 3.2 F_1 on the adversarial set respectively.

Methods	**OntoNotes5**				**Lap14**	
	Ori	Adv. full	Adv. entity	Adv. context	Ori.	Adv.
Tagging	89.8	**56.6**	**61.9**	83.6	62.3	44.5
MRC	90.0	55.3	61.3	83.3	63.2	46.9
PeerDA	**90.1**	55.9	61.0	**84.1**	**65.9**	**50.1**

Table 10: Robustness experiments against adversarial attacks. The results are reported on both original (Ori.) sets and the adversarial (Adv.) sets.

Methods	OntoNotes5	Lap14	CUAD	Social21
MRC+MenReplace	91.1	63.7	45.2	50.8
PeerDA	**92.4**	**65.9**	**52.3**	**58.1**

Table 11: Performance on peer-driven DA approaches.

A.5 Peer-driven DA:

We compare PeerDA with Mention Replacement (MenReplace) (Dai and Adel, 2020), another Peer-driven DA approach randomly replaces a span mention in the context with another mention of the same category in the training set. The results of four SpanID tasks are presented in Table 11. PeerDA exhibits better performance than MenReplace on all four tasks. In addition, MenReplace would easily break the text coherence as a result of putting span mentions into the incompatible context, while PeerDA can do a more natural augmentation without harming the context.

A.6 Error Analysis:

In order to know the typical failure of PeerDA, we randomly sample 100 error cases from **Ontonotes5** test set for analysis. As shown in Table 12, there are four major groups:

- *Multiple Labels*: PeerDA would assign multiple labels to the same detected span. And in most cases (35/41), this error occurs among similar categories, such as `LOC`, `GPE`, and `ORG`.
- *Incorrect Label*: Although spans are correctly detected, PeerDA assigns them the wrong categories. Note that MRC even cannot detect many of those spans (23/37). As a result, PeerDA significantly improves the model's capability to detect spans, but still faces challenges in category classification.
- *Missing Prediction*: Compared to MRC, PeerDA tends to predict more spans. Therefore it alleviates the missing prediction issue that MRC mostly suffers.
- *Other Errors*: There are several other errors, such as the incorrect span boundary caused by articles or nested entities.

Multiple Labels	*I'm in Atlanta.* **Gold**: ("Atlanta", GPE) **PeerDA**: ("Atlanta", GPE); ("Atlanta", LOC) (41%) **MRC**: ("Atlanta", GPE) ("Atlanta", LOC) (3%)
Incorrect Label	*Why did it take us to get Sixty Minutes to do basic reporting to verify facts?* **Gold**: ("Sixty Minutes", ORG) **PeerDA**: ("Sixty Minutes", WORK_OF_ART) (37%) **MRC**: ("Sixty Minutes", WORK_OF_ART) (20%)
Missing Prediction	*Coming to a retailer near you, PlayStation pandemonium.* **Gold**: ("PlayStation", PRODUCT) **PeerDA**: $\emptyset$ (19%) **MRC**: $\emptyset$ (74%)
Other Errors	*I was guarded uh by the British Royal Marines actually because unfortunately they've had now um uh roadside bombs down there not suicide bombs.* **Gold**: ("the British Royal Marines",ORG) **PeerDA**: ("Royal Marines",ORG) (3%) **MRC**: ("Royal Marines",ORG) (3%)

Table 12: Error analysis of base-sized PeerDA and MRC models on **Ontonotes5** test set. We randomly select 100 examples from the test set and compare the predictions and error percentage of the two models.