Out-of-Domain Discourse Dependency Parsing via Bootstrapping: An Empirical Analysis on Its Effectiveness and Limitation

Noriki Nishida and **Yuji Matsumoto**
RIKEN Center for Advanced Intelligence Project, Japan
{noriki.nishida, yuji.matsumoto}@riken.jp

Abstract

Discourse parsing has been studied for decades. However, it still remains challenging to utilize discourse parsing for real-world applications because the parsing accuracy degrades significantly on out-of-domain text. In this paper, we report and discuss the effectiveness and limitations of bootstrapping methods for adapting modern BERT-based discourse dependency parsers to out-of-domain text without relying on additional human supervision. Specifically, we investigate self-training, co-training, tri-training, and asymmetric tri-training of graph-based and transition-based discourse dependency parsing models, as well as confidence measures and sample selection criteria in two adaptation scenarios: monologue adaptation between scientific disciplines and dialogue genre adaptation. We also release COVID-19 Discourse Dependency Treebank (COVID19-DTB), a new manually annotated resource for discourse dependency parsing of biomedical paper abstracts. The experimental results show that bootstrapping is significantly and consistently effective for unsupervised domain adaptation of discourse dependency parsing, but the low coverage of accurately predicted pseudo labels is a bottleneck for further improvement. We show that active learning can mitigate this limitation.

1 Introduction

Discourse parsing aims to uncover structural organization of text, which is useful in Natural Language Processing (NLP) applications such as document summarization (Louis et al., 2010; Hirao et al., 2013; Yoshida et al., 2014; Bhatia et al., 2015; Durrett et al., 2016; Xu et al., 2020), text categorization (Ji and Smith, 2017; Ferracane et al., 2017), question answering (Verberne et al., 2007; Jansen et al., 2014), and information extraction (Quirk and Poon, 2017). In particular, dependency-style representation of discourse structure has been studied intensively in recent years (Asher and Lascarides, 2003; Hirao et al., 2013; Li et al., 2014b; Morey et al., 2018; Hu et al., 2019; Shi and Huang, 2019). Figure 1 shows an example of discourse dependency structure, which is recorded in COVID-19 Discourse Dependency Treebank (COVID19-DTB), a new manually annotated resource for discourse dependency parsing of biomedical abstracts.

State-of-the-art discourse dependency parsers are generally trained on a manually annotated treebank, which is available in a limited number of domains, such as RST-DT (Carlson et al., 2001) for news articles, SciDTB (Yang and Li, 2018) for NLP abstracts, and STAC (Asher et al., 2016) and Molweni (Li et al., 2020) for multi-party dialogues. However, when the parser is applied directly to out-of-domain documents, the parsing accuracy degrades significantly due to the domain shift problem. In fact, we normally face this issue in the real world because human supervision is generally scarce and expensive to obtain in the domain of interest.

Unsupervised Domain Adaptation (UDA) aims to adapt a model trained on a source domain, where a limited amount of labeled data is available, to a target domain, where only unlabeled data is available. *Bootstrapping* (or pseudo labeling) has been shown to be effective for the UDA problem of syntactic parsing (Steedman et al., 2003b,a; Reichart and Rappoport, 2007; Søaard and Rishøj, 2010; Weiss et al., 2015). In bootstrapping for syntactic parsing, we first train a model on the labeled source sentences, the model is used to give *pseudo labels* (i.e., parse trees) to unlabeled target sentences, and then the model is retrained on the manually and automatically labeled sentences.

On the contrary, despite the significant progress achieved in discourse parsing so far (Li et al., 2014b; Ji and Eisenstein, 2014; Joty et al., 2015; Perret et al., 2016; Wang et al., 2017; Kobayashi

Transactions of the Association for Computational Linguistics, vol. 10, pp. 127–144, 2022. https://doi.org/10.1162/tacl_a_00451
Action Editor: Shay Cohen. Submission batch: 6/2021; Revision batch: 9/2021; Published 2/2022.

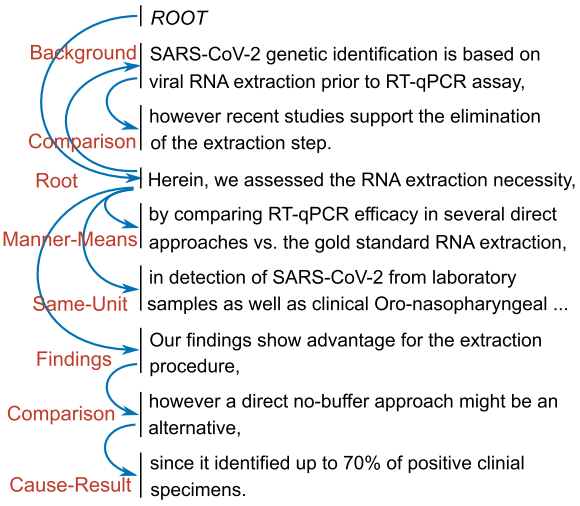


Figure 1: An example of discourse dependency structure for a COVID-19 related biomedical paper abstract (Israeli et al., 2020), which we manually annotated for our new dataset.

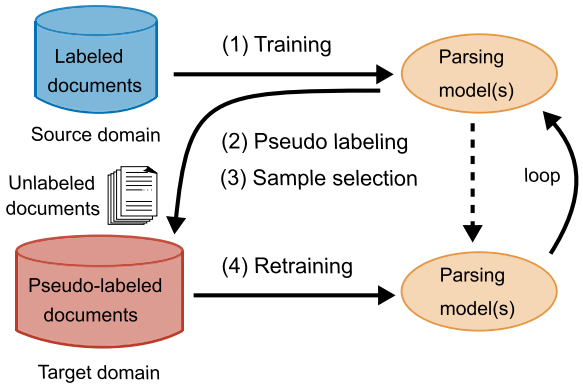


Figure 2: An overview of our bootstrapping system for unsupervised domain adaptation of discourse dependency parsing.

et al., 2020; Koto et al., 2021), bootstrapping for the UDA problem of discourse parsing is still not well understood. Jiang et al. (2016) and Kobayashi et al. (2021) explored how to enrich the labeled dataset using bootstrapping methods; however, their studies are limited to the *in-domain* setup, where the labeled and unlabeled datasets are derived from the *same* domain. In contrast to these studies, we focus on the more realistic and challenging scenario, namely, *out-of-domain* discourse parsing, where the quality and diversity of the pseudo-labeled dataset become more crucial for performance enhancement.

In this paper, we perform a series of analyses of various bootstrapping methods in UDA of modern BERT-based discourse dependency parsers and report the effectiveness and limitations of these approaches. Figure 2 shows an overview of our bootstrapping system. Specifically, we investigate self-training (Yarowsky, 1995), co-training (Blum and Mitchell, 1998; Zhou and Goldman, 2004), tri-training (Zhou and Li, 2005), and asymmetric tri-training (Saito et al., 2017) of graph-based and transition-based discourse dependency parsing models, as well as confidence measures and sample selection criteria in two adaptation scenarios: monologue adaptation between scientific disciplines and dialogue genre adaptation. We show that bootstrapping improves out-of-domain discourse dependency parsing significantly and consistently across different adaptation setups.

Our analyses also reveal that bootstrapping has a difficulty in creating pseudo-labeled data that is both diverse and accurate, which is a current limiting factor in further improving accuracy, and furthermore it is difficult to boost the coverage by simply increasing the number of unlabeled documents. We show that an *active learning* approach can be an effective solution to the limitation.[1]

The rest of this paper is organized as follows: Section 2 provides an overview of related studies. Section 3 clarifies the problem and describes the methodology: bootstrapping algorithms, discourse dependency parsing models, confidence measures, and sample selection criteria. Section 4 describes the details of COVID19-DTB. Section 5 describes the experimental setup. In Section 6 we report and discuss the experimental results and provide practical recommendations for out-of-domain discourse dependency parsing. Finally, Section 7 concludes the paper.

2 Related Work

Various discourse parsing models have been proposed in the past decades. For constituency-style discourse structure like RST (Mann and Thompson, 1988), the parsing models can be categorized into the chart-based approach (Joty et al., 2013; Joty et al., 2015; Li et al., 2014a, 2016a), which finds the globally optimal tree using an efficient algorithm like dynamic programming, or the

[1]Our code is available at `https://github.com/norikinishida/discourse-parsing`. The COVID19-DTB dataset is also available at `https://github.com/norikinishida/biomedical-discourse-treebanks`.

transition-based (or sequential) approach (Marcu, 1999; Sagae, 2009; Hernault et al., 2010b; Feng and Hirst, 2014; Ji and Eisenstein, 2014; Wang et al., 2017; Kobayashi et al., 2020; Zhang et al., 2020; Koto et al., 2021), which builds a tree incrementally by performing a series of decisions. For dependency-style discourse structure like the RST variants (Hirao et al., 2013; Li et al., 2014b; Morey et al., 2018) or Segmented Discourse Representation Theory (Asher and Lascarides, 2003), the models can also be categorized into the graph-based approach (Li et al., 2014b; Yoshida et al., 2014; Afantenos et al., 2015; Perret et al., 2016) or the transition-based (sequential) approach (Muller et al., 2012; Hu et al., 2019; Shi and Huang, 2019). Recently, pre-trained transformer encoders such as BERT (Devlin et al., 2019) and SpanBERT (Joshi et al., 2019) have been shown to greatly improve discourse parsing accuracy (Guz and Carenini, 2020; Koto et al., 2021). In this paper, we are not aiming at developing novel parsing models. Instead, we aim to investigate the effectiveness and limitations of bootstrapping methods for adapting the modern BERT-based discourse parsers.

Manually annotated discourse treebanks are significantly scarce, and their domains are limited. For example, the most popular discourse treebank, RST-DT (Carlson et al., 2001), contains only 385 labeled documents in total. To address the lack of large-scale labeled data, a number of semi-supervised, weakly supervised, and unsupervised techniques have been proposed in the discourse parsing literature. Hernault et al. (2010a) proposed a semi-supervised method that utilizes unlabeled documents to expand feature vectors in SVM classifiers in order to achieve better generalization for infrequent discourse relations. Liu and Lapata (2018) and Huber and Carenini (2019) proposed to exploit document-level class labels (e.g., sentiment) as distant supervision to induce discourse dependency structures from neural attention weights. Badene et al. (2019a,b) investigated a data programming paradigm (Ratner et al., 2016), which uses rule-based labeling functions to automatically annotate unlabeled documents and trains a generative model on the weakly supervised data. Kobayashi et al. (2019) and Nishida and Nakayama (2020) proposed fully unsupervised discourse constituency parsers, which can produce only tree skeletons and rely strongly on pre-trained word embeddings or human prior knowledge on document structure.

Technically most similar to our work, Jiang et al. (2016) and Kobayashi et al. (2021) proposed to enlarge the training dataset using a combination of multiple parsing models. Jiang et al. (2016) used co-training for enlarging the RST-DT training set with 2,000 *Wall Street Journal* articles, with a focus on improving classification accuracy on infrequent discourse relations. Kobayashi et al. (2021) proposed to exploit discourse subtrees that are agreed by two different models for enlarging the RST-DT training set. Interestingly, their proposed methods improved the classification accuracy especially for infrequent discourse relations.

These studies mainly assume the in-domain scenario and focus on enlarging the labeled set (e.g., RST-DT training set) using in-domain unlabeled documents, and the system evaluation is generally performed on the same domain with the original labeled set (e.g., RST-DT test set). In this paper, instead, we particularly focus on the UDA scenario, where the goal is to parse the target-domain documents accurately without relying on human supervision in the target domain. We believe this research direction is important for developing usable discourse parsers, because a target domain to which one would like to apply a discourse parser is normally different from the domains/genres of existing corpora, and manually annotated resources are rarely available in most domains/genres.

3 Method

3.1 Problem Formulation

The input is a document represented as a sequence of clause-level (in single-authored text) or utterance-level (in multi-party dialogues) spans called Elementary Discourse Units (EDUs).[2] Our goal is to derive a discourse dependency structure, $y = \{(h, d, r) \mid 0 \leq h \leq n, 1 \leq d \leq n, r \in \mathcal{R}\}$, given the input EDUs, $x = e_0, e_1, \ldots, e_n$, which is analogous to syntactic dependency structure. A discourse dependency, (h, d, r), represents that the d-th EDU (called *dependent*) relates to the h-th EDU (called *head*) directly with the discourse relation $r \in \mathcal{R}$. Each EDU except for the root node, e_0, has a single head.

In this paper, we assume that we have a limited number of labeled documents in the source

[2]We call both single-authored text and multi-party dialogues as *documents*.

domain, while a large collection of unlabeled documents is available in the target domain. In particular, we assume that the source and target domains have different data distributions lexically or rhetorically (e.g., vocabulary, document length, and discourse relation distributions), but the domains share the same annotation scheme (e.g., definition of discourse relation classes). Our task is to adapt a parsing model (or models) trained in the source domain to the target domain using the unlabeled target data.

3.2 Bootstrapping

The aim of this paper is to investigate the effectiveness and limitations of various bootstrapping methods in UDA of modern BERT-based discourse dependency parsers. We show the overall flow of the bootstrapping methods in Figure 2. Initially we have a small set of labeled documents, $\mathcal{L}^s$, in the source domain, and a large collection of unlabeled documents, $\mathcal{U}^t$, in the target domain. Then the bootstrapping procedure works as follows:

(1) Train initial models on $\mathcal{L}^s = \{(x^s, y^s)\}$.

(2) Parse unlabeled documents $x^t \in \mathcal{U}^t$ using the current model f, such as, $f: \mathcal{U}^t \rightarrow \mathcal{L}^t = \{(x^t, f(x^t))\}$.[3,4]

(3) Measure the confidence scores of the pseudo-labeled data and select a subset, $\tilde{\mathcal{L}}^t \subset \mathcal{L}^t$, that is expected to be reliable and useful.

(4) Retrain the models on $\mathcal{L}^s \cup \tilde{\mathcal{L}}^t$ for several epochs (set to 3 in this work).

Steps (2)-(4) loop for many rounds until a predefined stopping criterion is met.

Bootstrapping can be interpreted as a methodology where *teachers* generate pseudo supervision for *students*, and the students learn the task on it. Existing bootstrapping methods vary depending on how the teacher and student models are used. In this paper, we specifically explore the following bootstrapping methods: self-training (Yarowsky, 1995; McClosky et al., 2006; Reichart and Rappoport, 2007; Suzuki and Isozaki, 2008; Huang and Harper, 2009), co-training (Blum and Mitchell, 1998; Zhou and Goldman, 2004; Steedman et al., 2003b,a), tri-training (Zhou and Li, 2005; Weiss et al., 2015; Ruder and Plank, 2018), and asymmetric tri-training (Saito et al., 2017).

Self-Training Self-Training (ST) starts with a single model f trained on $\mathcal{L}^s$. The overall procedure is the same as the one described above. The single model is both a *teacher* and a *student* for itself. Thus, it is difficult for the model to obtain novel knowledge (or supervision) that the model has not learn, and its errors may be amplified by the retraining cycle.

Co-Training Co-Training (CT) starts with two parsing models, f_1 and f_2, that are expected to have different *inductive biases* with each other. The two models are pre-trained on the same $\mathcal{L}^s$. In Step 2, each model independently parses the unlabeled documents: $\mathcal{U}^t \rightarrow \mathcal{L}_i^t$ $(i = 1, 2)$. In Step 3, each of the pseudo-labeled sets are filtered by a selection criterion: $\mathcal{L}_i^t \rightarrow \tilde{\mathcal{L}}_i^t$. In Step 4, each model f_i is retrained on $\mathcal{L}^s \cup \mathcal{L}_j^t$ $(j \neq i)$.

In CT, the two models teach each other. Thus, each model is the *teacher* and the *student* for the other model simultaneously. In contrast to ST, each model can obtain knowledge that it has not yet learned. CT can be viewed as enhancing the agreement between the models.

Tri-Training (TT) Tri-Training (TT) consists of three different models, f_1, f_2, and f_3, which are initially trained on the same $\mathcal{L}^s$. In contrast to CT, where the single teacher f_i is used to generate pseudo labels $\tilde{\mathcal{L}}_i^t$ for the student f_j $(j \neq i)$, TT uses two teachers, f_i and f_j $(j \neq i)$, to generate a pseudo-labeled set $\mathcal{L}_{i,j}^t$ for the remaining student f_k $(k \neq i, j)$. We measure the confidence for the pair of teachers' parse trees, (y_i^t, y_j^t), using the ratio of *agreed* dependencies (described in Subsection 3.4), based on which we determine whether or not to include the teachers' predictions in the pseudo-labeled set.

Asymmetric Tri-Training (AT) Asymmetric Tri-training (AT) is an extension of TT for UDA. A special domain-specific model f_1^t is used only

[3]For bootstrapping methods that employ multiple models (e.g., co-training), $\mathcal{L}^t$ is created for each model f.

[4]In our experiments, for every bootstrapping round we used 5K sampled documents instead of the whole unlabeled documents $\mathcal{U}^t$, because parsing the whole documents at every bootstrapping round is computationally expensive and does not scale to a large-scale dataset. The 5K samples were flashed for every bootstrapping round.

for test inference; the other two models, f_2 and f_3, are used only to generate pseudo labels $\tilde{\mathcal{L}}^t$. The domain-specific model f_1^t is retrained on only $\tilde{\mathcal{L}}^t$, while f_2 and f_3 are retrained on $\mathcal{L}^s \cup \tilde{\mathcal{L}}^t$.

3.3 Parsing Models

We employ three types of BERT-based discourse dependency parsers: (1) A graph-based arc-factored model (McDonald et al., 2005) with a biaffine attention mechanism (Dozat and Manning, 2017), (2) a transition-based shift-reduce model (Nivre, 2004; Chen and Manning, 2014; Kiperwasser and Goldberg, 2016), and (3) the backward variant of the shift-reduce model.

EDU Embedding We compute EDU embeddings using a pre-trained Transformer encoder. This manner is common across the three parsing models, though the Transformer parameters are untied and fine-tuned separately. Specifically, we first break down the input document into non-overlapping segments of 512 subtokens, and then encode each segment independently by the Transformer encoder. Lastly, we compute EDU-level span embeddings as a concatenation of the Transformer output states at the span endpoints (w_i and w_j) and the span-level syntactic head word[5] w_k, i.e., $[\boldsymbol{w}_i; \boldsymbol{w}_j; \boldsymbol{w}_k]$.

Arc-Factored Model Arc-Factored Model (A) is a graph-based dependency parser, which can find the globally optimal dependency structure using dynamic programming. Specifically, we employ the biaffine attention model (Dozat and Manning, 2017) for computing dependency scores $s(h, d) \in \mathbb{R}$, and we decode the optimal structure y^* using Eisner Algorithm, such that the tree score $\sum_{(h,d)\in y} s(h, d)$ is maximized. We predict the discourse relation classes for each unlabeled dependency $(h, d) \in y^*$ using another biaffine attention layer and MLP, namely, $r^* = \operatorname{argmax}_r P(r \mid h, d)$. To reduce the computational time for inference, we employed the Hierarchical Eisner Algorithm (Zhang et al., 2021), which decodes dependency trees from the sentence level to the paragraph level and then to the whole text level.

[5] A span-level syntactic head word is a token whose parent in the syntactic dependency graph is ROOT or is not within the EDU's span. When there are multiple head words in an EDU, we choose the left most one. We used the spaCy *en_core_web_sm* model to obtain the syntactic dependency graph.

Shift-Reduce Model Shift-Reduce Model (S) is a transition-based dependency parser, which builds a dependency structure incrementally by executing a series of local *actions*. Specifically, we employ the arc-standard system proposed by Nivre (2004), which has a *buffer* to store the input EDUs to be analyzed and a *stack* to store the in-progress subtrees and defines the following action classes: SHIFT, RIGHT-ARC-l, and LEFT-ARC-l. We decode the dependency structure y^* using a greedy search algorithm, i.e., taking the action a^* that is valid and the most probable at each decision step: $a^* = \operatorname{argmax}_a P(a \mid \sigma)$, where σ denotes the parsing configuration.

Backward Shift-Reduce Model We expect that different inductive biases can be introduced by processing the document from the back. As the third model option, we develop a backward variant of the Shift-Reduce Model (B), which processes the input sequence in the reverse order.

3.4 Confidence Measures

The key challenge in bootstrapping on out-of-domain data is how to assess the reliability (or usefulness) of the pseudo labels and how to select an error-free and high-coverage subset. We define *confidence measures* to assess the reliability of the pseudo-labeled data. In Section 3.5, we define *selection criteria* to filter out unreliable pseudo-labeled data based on their confidence scores.

Model-based Confidence For the bootstrapping methods that use a single teacher to generate a pseudo-labeled set (i.e., ST, CT), we define the confidence of the teacher model based on predictive probabilities of the decisions used to build a parse tree. A discourse dependency structure consists of a set (or series) of decisions. Therefore, we use the average of the predictive probabilities over the decisions.[6] How to calculate

[6] We also tested a model-based confidence measure using the entropy of predictive probabilities, where we replaced the predictive probability of a decision (e.g., $P(h^* \mid d)$) with the corresponding (negative) entropy, e.g., $-H(h \mid d) = \sum_{0\le h\le n} P(h \mid d) \log P(h \mid d)$. Entropy has been used especially in the active learning literature to calculate data uncertainty (Li et al., 2016b; Kasai et al., 2019). However, the predictive probabilities outperformed the entropy counterparts consistently in our experiments. Thus, we adopted the predictive probabilities for the model-based confidence measure.

the model-based confidence measure $C(x, y)$ depends on the parsing models:

- Arc-Factored Model:

$$C(x,y) = \frac{1}{2n}\sum_{d=1}^{n}\{P(h \mid d) + P(r \mid h, d)\},$$

where $(h, d, r) \in y$.

- Shift-Reduce Model, Backward Model:

$$C(x,y) = \frac{1}{|A(x,y)|}\sum_{(a,\sigma)\in A(x,y)} P(a \mid \sigma),$$

where $A(x, y)$ denotes the action and configuration sequence to produce the parse tree y for x.

Agreement-based Confidence For the bootstrapping methods that use multiple teachers to generate a pseudo-labeled set (i.e., TT, AT), we use the agreement level between the two teacher models as the confidence for the pseudo-labeled data. Specifically, we compute the rate of labeled dependencies agreed between two predicted structures, y_i and y_j, as follows:

$$C(x, y_i, y_j) = \frac{1}{n}\sum_{d=1}^{n} \mathbb{1}\left[h_d^i = h_d^j \wedge r_d^i = r_d^j\right],$$

where $\mathbb{1}[\cdot]$ is the indicator function, and h_d^i and r_d^i denote the head and the discourse relation class for the dependent d in y_i, respectively. It is worth noting that both y_i and y_j have the same number of dependencies, n. The higher the percentage is, the more correct dependencies are expected to be included.

3.5 Sample Selection Criteria

Inspired by Steedman et al. (2003a), we define two kinds of sample selection criteria, each of which focuses on the reliability (i.e., accuracy) and the usefulness (i.e., training utility) of the data, respectively.

Rank-above-k This is a reliability-oriented selection criterion. We keep only the top $N \times k$ samples with higher confidence scores, where N is the number of candidate pseudo-labeled data, and $k \in [0.0, 1.0]$. Specifically, we first rank the candidate pseudo-labeled data based on the teacher-side confidence scores, and then we select a subset that satisfies $\mathrm{R}(x) \leq N \times k$. where $\mathrm{R}(x) \in [1, N]$ denotes the ranking of x.

COVID19-DTB	**SciDTB**
Root	Root
Elaboration	Elaboration, Progression, Summary
Comparison	Contrast, Comparison
Cause-Result	Cause-Effect, Explain
Condition	Condition
Temporal	Temporal
Joint	Joint
Enablement	Enablement
Manner-Means	Manner-Means
Attribution	Attribution
Background	Background
Findings	Evaluation
Textual-Organization	–
Same-Unit	Same-Unit

Table 1: Discourse relation classes in COVID19-DTB and their correspondences with SciDTB's classes.

Rank-diff-k This is a utility-oriented selection criterion. In contrast to Rank-above-k, which relies only on the teacher-side confidence, this criterion utilizes both the teacher-side and the student-side confidence scores. This criterion retain the pseudo-labeled data whose relative ranking on the teacher side is higher than the relative ranking on the student side by a margin k or more. Specifically, after ranking the candidates independently for each side, we compute the gap of the relative rankings on the two sides, and then select a subset that meets $\mathrm{R}_{\text{teacher}}(x) + k \leq \mathrm{R}_{\text{student}}(x)$.

4 COVID19-DTB

We release a new discourse dependency treebank for scholarly paper abstracts on COVID-19 and related coronaviruses like SARS and MERS in order to test unsupervised domain adaptation of discourse dependency parsing. We name our new treebank COVID-19 Discourse Dependency Treebank (COVID19-DTB).

4.1 Construction

We followed the RST-DT annotation guideline (Carlson and Marcu, 2001) for EDU segmentation. Based on SciDTB and Penn Discourse Treebank (PDTB) (Prasad et al., 2008), we defined 14 discourse relation classes shown in Table 1. We carefully analyzed the annotation data of SciDTB and found that some classes are hard to discriminate even for humans, which can lead

	COVID19-DTB	SciDTB
Total number of documents	300	1045 (unique: 798)
Total number of EDUs	6005	15723
Avg number of EDUs / doc	20.0	15.0
Avg dependency distance	2.7	2.5
Max. dependency distance	38	26
Avg Root position	6.6	3.9

Table 2: Dataset statistics for the COVID19-DTB and SciDTB datasets.

to undesirable inconsistencies in the new dataset. Thus, we have merged some classes, such as Cause-Effect + Explain → Cause-Result. Some classes are also renamed from SciDTB to fit the biomedical domain, such as Evaluation → Findings.

First, we sampled 300 abstracts randomly from the 2020 September snapshot of The COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020), which contains over 500,000 scholarly articles on COVID-19 and related coronaviruses like SARS and MERS. Then, the 300 abstracts were segmented into EDUs manually by the authors. Then, we employed two professional annotators to give gold discourse dependency structures to the 300 abstracts. The annotators were trained using a few examples and a manual guideline, and then they annotated the 300 abstracts independently.[7] We divided the results into development and test splits, each of which consists of 150 examples.

4.2 Corpus Statistics

Table 2 and Figure 3 show the statistics and the discourse relation distribution of COVID19-DTB. We also show the statistics and the distribution of SciDTB for comparison. We mapped discourse relations in SciDTB to the corresponding classes in COVID19-DTB. We removed the Root relations in computing the proportions.

The average number of EDUs per document in each corpus was 20.0 and 15.0, respectively. Although the average dependency distances in the two corpora are almost the same (2.7 vs. 2.5), the maximum dependency distance of COVID19-DTB is significantly longer than that of SciDTB. Furthermore, the average position of Root's direct dependent is located further back in COVID19-DTB (6.6 vs. 3.9). Although the overall discourse

[7]The inter-annotator agreement is thus not calculated in the current version of the dataset. Instead, we had several discussions with each annotator to maintain the annotation consistency at a satisfactory level.

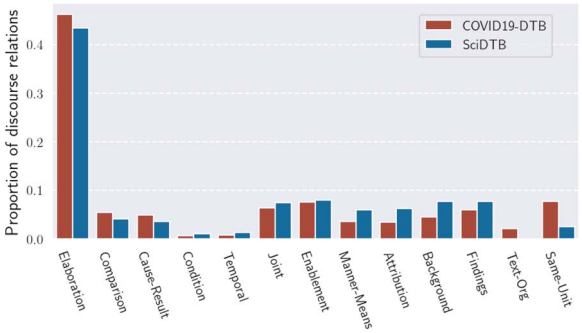


Figure 3: Distributions of discourse relation classes in COVID19-DTB and SciDTB. Discourse relations in SciDTB are mapped to the corresponding classes in COVID19-DTB.

relation distributions look similar, the proportions of Elaboration and Same-Unit are larger in COVID19-DTB. These differences reflect the fact that biomedical abstracts tend to be longer, have more complex sentences with embedded clauses, and contain more detailed information, suggesting the difficulty of discourse parser adaptation across the two domains.

5 Experimental Setup

Datasets We evaluated the bootstrapping methods on two UDA scenarios: The first setup was a monologue adaptation between scientific disciplines: NLP and biomedicine (especially on COVID-19), which is actually an important scenario because there is still no text-level discourse treebank on biomedical documents. We used the training split of **SciDTB** (Yang and Li, 2018) as the labeled source dataset, which contains 742 manual discourse dependency structures on the abstracts in ACL Anthology. We also used the 2020 September snapshot of **CORD-19** (Wang et al., 2020) as the unlabeled target dataset, which contains about 76,000 biomedical abstracts. We used the development and test splits of **COVID19-DTB** for validation and testing, respectively. The discourse relation labels in the SciDTB training set were mapped to the corresponding classes of COVID19-DTB. We mapped Textual-Organization relations in COVID19-DTB to Elaboration, because there is no corresponding class in SciDTB. We also mapped Temporal relations in the two datasets to Condition to reduce the significant class imbalance.

The second setup was an adaptation across dialogue genres, that is, dialogues in a multi-party

game and dialogues in Ubuntu Forum. We used the training split of **STAC** (Asher et al., 2016) as the labeled source dataset, which contains 887 manually labeled discourse dependency structures on multi-party dialogues in the game, *The Settlers of Catan*. We also used the **Ubuntu Dialogue Corpus** (Lowe et al., 2015) as the unlabeled target dataset, which contains dialogues extracted from the Ubuntu chat logs. We retained dialogues with 7-16 utterances and 2-9 speakers. We also removed dialogues with long utterances (more than 20 words). Finally, we obtained approximately 70,000 dialogues. We used the development and test splits of **Molweni** (Li et al., 2020) for validation and testing. Each split contains 500 manually labeled discourse dependency structures on multi-party dialogues derived from the Ubuntu Dialogue Corpus.

The unlabeled target documents in both setups were segmented into EDUs using a publicly available EDU segmentation tool (Wang et al., 2018).

Evaluation We employed the traditional evaluation metrics in dependency parsing literature, namely, Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS). We also used Root Accuracy (RA), which indicates how well a system can identify the most representative EDU in the document (i.e., the dependent of the special root node).

Implementation Details As the pre-trained transformer encoders, we used SciBERT (Beltagy et al., 2019) and SpanBERT (Joshi et al., 2019) in the first and second adaptation setups, respectively. The dimensionality of the MLPs in the arc-factored model and the shift-reduce models are 100 and 128, respectively. We used AdamW and Adam optimizers for optimizing the transformer's parameters (θ_{bert}) and the task-specific parameters (θ_{task}), respectively, following Joshi et al. (2019). We first trained the base models on the labeled source dataset using the following hyper-parameters: batch size = 1, learning rate (LR) for $\theta_{\text{bert}} = 2\text{e}^{-5}$, LR for $\theta_{\text{task}} = 1\text{e}^{-4}$, warmup steps = 2.4K. Then, we ran the bootstrapping methods using the models with: batch size = 1, LR for $\theta_{\text{bert}} = 2\text{e}^{-6}$, LR for $\theta_{\text{task}} = 1\text{e}^{-5}$, warmup steps = 7K. We trained all approaches for a maximum of 40 epochs. We applied early stopping when the validation LAS does not increase for 10 epochs.

		Abstracts			Dialogues	
Method	**Selection**	**LAS**	**UAS**	**RA**	**LAS**	**UAS**
Source-only (A)	–	61.3	74.8	82.0	29.9	55.1
Source-only (S)	–	61.8	74.5	78.0	33.2	66.1
Source-only (B)	–	60.0	72.9	78.0	29.2	55.6
ST (A ← A)	above-0.6	65.8	78.7	**88.7**	34.7	60.6
ST (S ← S)	above-0.6	65.3	76.9	84.7	37.9	67.4
CT (A ← S)	above-0.6	**66.2**	78.1	86.0	38.0	64.8
CT (S ← A)	above-0.6	66.1	78.2	86.0	39.1	64.4
CT (A ← S)	diff-100	66.0	78.3	88.0	38.5	66.5
CT (S ← A)	diff-100	**66.2**	**78.8**	84.7	**39.5**	66.0
CT (S ← B)	above-0.6	65.3	76.8	84.0	38.1	67.2
CT (B ← S)	above-0.6	65.6	76.9	87.3	38.5	67.4
CT (S ← B)	diff-100	65.5	76.8	86.0	39.1	67.5
CT (B ← S)	diff-100	65.5	76.6	86.7	39.2	**67.7**
TT (A ← S, B)	above-0.6	65.9	78.5	87.3	38.5	66.6
TT (S ← A, B)	above-0.6	65.9	78.4	86.0	39.1	66.7
TT (A ← S, B)	diff-100	65.4	77.4	86.7	38.6	66.8
TT (S ← A, B)	diff-100	65.1	77.7	87.3	38.9	66.5
AT (A ← S, B)	above-0.6	64.9	77.3	85.3	36.9	66.7
AT (S ← A, B)	above-0.6	65.3	77.4	**88.7**	38.6	63.2
AT (A ← S, B)	diff-100	65.3	77.6	84.7	36.9	65.7
AT (S ← A, B)	diff-100	64.6	77.6	85.3	38.2	61.9

Table 3: LAS for methods with and without bootstrapping in the two UDA setups. Arrows indicate the teacher and student models: For example, TT (S ← A, B) shows the test performance of the Shift-Reduce Model (S) that is trained with the Arc-Factored Model (A) and the Backward Shift-Reduce Model (B) using Tri-Training (TT). RA is omitted for the dialogue adaptation setup because the accuracy is nearly 100% for most systems.

6 Results and Discussion

6.1 Effectiveness

We verified the effectiveness of bootstrapping methods on the two UDA scenarios. We evaluated the source-only models, which were trained only on the labeled source dataset, as the baseline. Table 3 shows the results. The bootstrapping methods consistently gave gains in performance regardless of the adaptation scenarios. The best systems were CT (A ← S) with Rank-above-0.6 and CT (S ← A) with Rank-diff-100, which outperformed the source-only systems (e.g., Source-only (S)) by more than 4.4 LAS points on the monologue setup (NLP → COVID-19) and by more than 6.3 LAS points on the dialogue setup (Game → Ubuntu Forum), respectively. CT, TT, and AT tended to achieve higher accuracy than ST, particularly in the dialogue adaptation setup. These results indicate that bootstrapping is significantly and consistently effective for UDA of discourse dependency parsing in various adaptation scenarios, and that employing multiple models

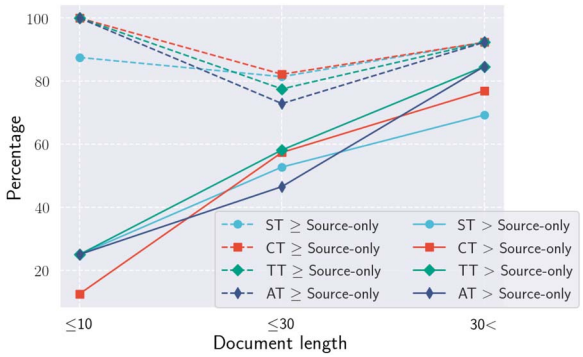


Figure 4: How bootstrapping methods improve performance as a function of document length (i.e., the number of EDUs) in the target domain.

is particularly effective in reducing the unintended tendency of ST to amplify its own errors.

Next, we further analyzed in what kind of documents the bootstrapping system is particularly effective. We divided the COVID19-DTB test set into bins by the number of EDUs in each document ($n \leq 15$, $15 < n \leq 30$, $n > 30$), and for each bin we examined the percentage of examples improved by the bootstrapping systems over the source-only system. Figure 4 shows the results. When the document length was 10 or shorter, there was no improvement in most examples; however, when the length was longer than 30, the percentage was jumped to around 80% with CT, TT, and AT. These results indicate that the longer the documents (or maybe the higher the document complexity) in the target domain, the greater the benefit of bootstrapping.

We also investigated the importance of employing different types of parsing models in CT. The theoretical importance of employing models with different views (or inductive biases) in CT has been discussed (Blum and Mitchell, 1998; Abney, 2002; Zhou and Goldman, 2004). We trained base models with the same neural architecture but with different initial parameters on the labeled source dataset, and then retrained them using CT. We can see from the results in Table 4 that the LAS of CT using different model types is consistently higher than that of CT with the same model types, suggesting empirically that it is effective to employ different model types in bootstrapping.

6.2 Analysis of Confidence Measures

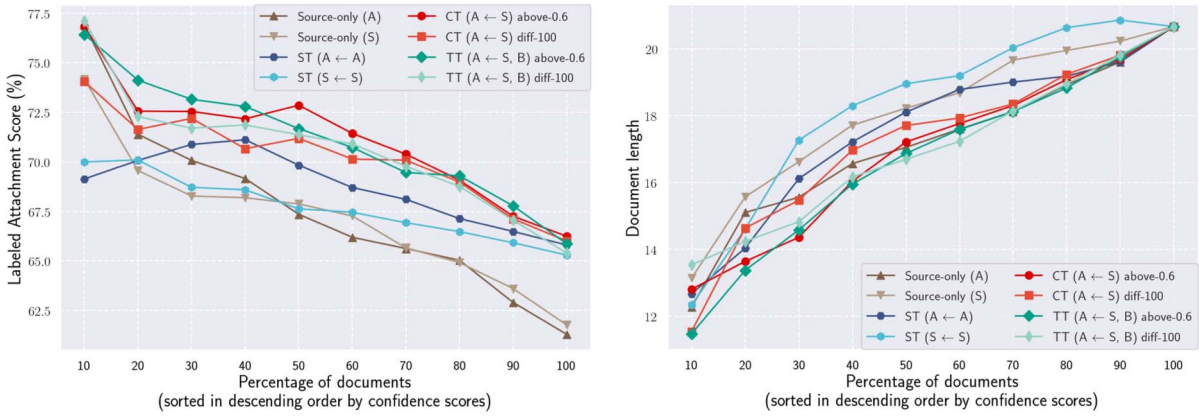
One of the key challenges in bootstrapping for UDA is to assess the true reliability of pseudo-labeled data. Here, we analyzed the confidence measures.

		Abstracts		
Method	**Selection**	**LAS**	**UAS**	**RA**
CT (A ← S)	above-0.6	66.2	78.1	86.0
CT (S ← A)	above-0.6	66.1	78.2	86.0
CT (A ← A)	above-0.6	65.5	77.8	76.0
CT (S ← S)	above-0.6	65.5	77.9	86.0

Table 4: Comparison of co-training systems employing different types or the same type of parsing models.

Confidence Scores Correlate with Quality Regardless of the selection criteria, data with high confidence scores tend to be selected. Figure 5 (a) shows the relationships between the confidence scores and the parsing quality (LAS) in the target domain. Specifically, we calculated the confidence scores of each example in the COVID19-DTB test set, sorted the examples in descending order of their confidence scores, and evaluated LAS for each of the top $k\%$ subset. We confirmed that the confidence scores were roughly correlated with the parsing quality, and the top candidates of higher confidence tended to be more accurate than the ones of lower confidence. For example, if we restricted the test data to the top 10% with the highest confidence scores, the LAS of CT (A ← S) with Rank-above-0.6 was 76.8%, which was much higher than the LAS of this system on the full test set (i.e., 66.2%).

Confident (Accurate) Pseudo Labels are Biased Next, we examined what kind of documents are assigned with higher confidence scores. Figure 5 (b) shows the relationships between the confidence scores and the document length (i.e., the number of EDUs). We found the strong correlation between them: Documents with higher confidence scores are biased to shorter documents. This bias did not depend on the confidence measures (model-based vs. agreement-based), the sample selection criteria, and even the presence of bootstrapping. Based on these results, we can conjecture that longer documents tend to be of poor quality (low confidence) and less likely to be included in the selected pseudo-labeled set $\tilde{\mathcal{L}}^t$. This conjecture further implies that the current bottleneck of the



(a) Confidence vs. LAS

(b) Confidence vs. Document length

Figure 5: Relationships between the confidence scores and the parsing quality (LAS) or document length. We used the examples in the COVID19-DTB test set. The examples are sorted in descending order of the confidence scores.

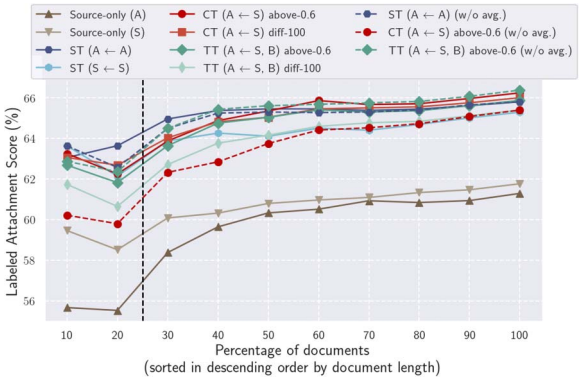


Figure 6: Relationships between the document length and the parsing quality (LAS). Documents are sorted in descending order of the document length.

bootstrapping systems is the low coverage of the selected pseudo-labeled set $\tilde{\mathcal{L}}^t$.

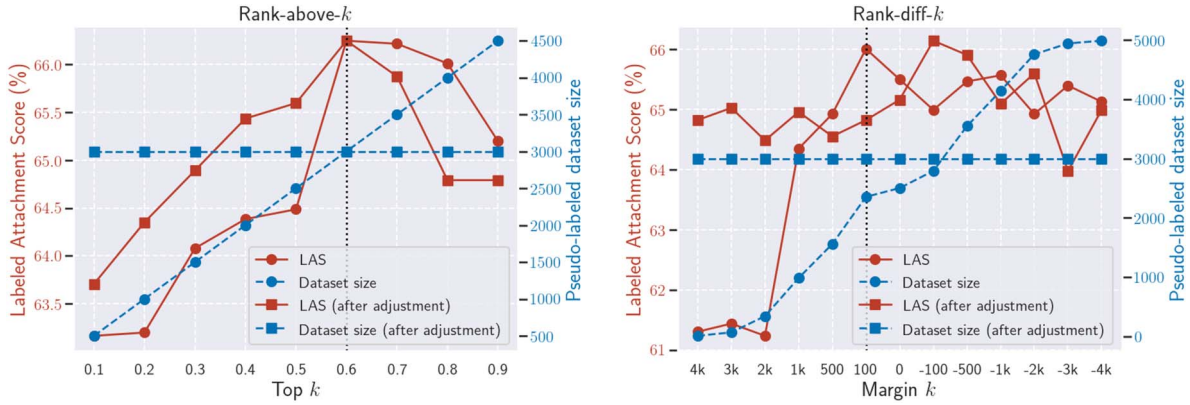
Low Coverage of Accurate Pseudo Labels Based on the above conjecture, it is natural to expect that there is too little accurate supervision for longer documents in the selected pseudo-labeled set, and that the parsing accuracy of the bootstrapping systems drop especially for longer documents. Figure 6 shows the relationships between the parsing quality and the document length. The use of bootstrapping methods improved the overall performance over the source-only systems; however, regardless of the bootstrapping types, the performance dropped significantly for longer documents. These results confirm the shortage of accurate supervision for longer documents in the selected pseudo-labeled set.

The Problem is Not Sampling, but Prediction To alleviate this low-coverage issue, we modified the confidence measures defined in Subsection 3.4 to select longer documents more aggressively for the pseudo-labeled set $\tilde{L}^t$. We simply omitted the averaging calculation over the decisions. However, as shown in Figure 6 (see the results with ‘‘w/o avg.’’), the performance degradation tendency for longer documents did not change. This fact indicates that the current bottleneck is not the low coverage of the *selected* pseudo-labeled set $\tilde{\mathcal{L}}^t$, but the low coverage of accurate supervision in the candidate pseudo-labeled data pool, that is, $\mathcal{L}^t$.

6.3 Analysis of Selection Criteria

Another important challenge in bootstrapping for UDA is to select an error-free and high-coverage subset from the candidate pseudo-labeled data pool. Here, we analyzed the selection criteria.

There Is a Reliability-Coverage Trade-off Varying the parameter k of Rank-above-k and Rank-diff-k, we examined the final parsing quality and the average number of selected pseudo-labeled data (out of 5K candidates). We trained and evaluated CT (A ← S) on the COVID19-DTB test set. Figure 7 (lines with circle markers) shows the results. Rank-above-k achieved a slightly higher performance than Rank-diff-k. However, Rank-diff-k achieved the best performance with less pseudo-labeled data. More interestingly, we confirmed that, for both criteria, there is a trade-off



(a) Rank-above-k vs. LAS, Selected data size

(b) Rank-diff-k vs. LAS, Selected data size

Figure 7: Impacts of the sample selection criteria (Rank-above-k, Rank-diff-k) for different parameters k. We also show the results when the number of selected pseudo-labeled data is adjusted to 3K.

between *reliability* (*precision*) and *coverage* (*recall*) in the selected pseudo-labeled set: When k was too strict (i.e., when k was too small for Rank-above-k, and when the margin k was too large for Rank-diff-k), the number of selected pseudo-labeled data was too small, resulting in lower LAS. When k was relaxed to some extent, the number of selected pseudo-labeled data increased and the LAS reached the highest LAS. However, if k was relaxed further, the accuracy decreased from the highest point due to the contamination of too much noisy supervision.

Quantity Is Not the Only Issue Next, we evaluated the sample selection criteria without the influence of the number of selected pseudo-labeled data. We selected the same number of pseudo-labeled data (set to 3K) across different k by adjusting the number of the unlabeled samples (set to 5K in the previous experiments) appropriately. For example, to select 3K pseudo-labeled data with Rank-above-0.2, we sampled 15K unlabeled data at each bootstrapping round. Figure 7 (lines with rectangle markers) shows the results. In the region where k was strict, the LAS curves improved compared to before the adjustment because the number of selected pseudo-labeled data increased. Meanwhile, in the region where k was relaxed, the LAS curves decreased or were retained because the selection size is decreased. More interestingly, even with this adjusted setting, the strictest k was not the best parameter. These results indicate that, although the number of selected pseudo-labeled data is important, it is not the only factor that determines the optimal parameter k, and that it is still difficult to identify truly useful pseudo-labeled data based on these sample selection criteria alone.

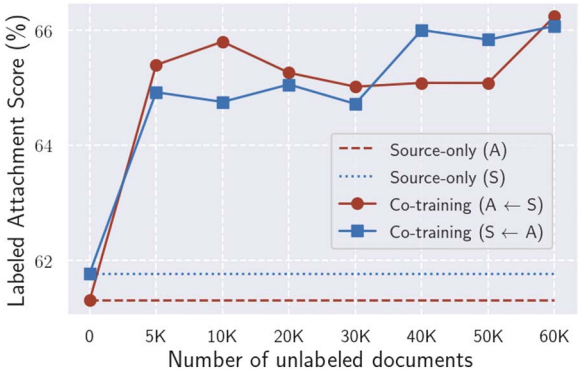


Figure 8: Impacts of the number of unlabeled target documents.

6.4 Increasing the Unlabeled Dataset Size

So far, we have demonstrated that the current major limitation of bootstrapping is the difficulty of generating *diverse* and *accurate* pseudo-labeled data pool $\mathcal{L}^t$. The most straightforward way for mitigating this low-coverage problem is to increase the number of unlabeled target documents. Figure 8 shows that increasing the number of unlabeled data with bootstrapping improved the parsing quality. However, the quality improvement saturated after 5K documents. These facts demonstrate that the low-coverage problem can not be mitigated by simply adding more unlabeled documents. We suspect this is because increasing the diversity of unlabeled documents does not always increase the diversity of accurately pseudo-labeled data.

Method	Confidence	LAS
Source-only (S)	–	33.2
CT (S ← A) w/ above-0.6	–	39.1
Source-only (S) + AL	random	45.2
Source-only (S) + AL	model-based	45.8
CT (S ← A) w/ above-0.6 + AL	random	45.9
CT (S ← A) w/ above-0.6 + AL	model-based	**46.4**
CT (S ← A) w/ above-0.6 + AL	agreement-based	46.3

Table 5: LAS for methods with and without active learning (AL).

6.5 Active Learning

A more direct and promising solution than increasing the unlabeled corpus size is to manually annotate a small amount of documents that the bootstrapping system can not analyze accurately. We tested the potential effectiveness of *active learning* (AL) (Settles, 2009). To emulate the AL process, we used the Molweni training set (9K dialogues) as the unlabeled target documents and leveraged the gold annotation. We first measured the confidence (or *uncertainty*) scores of each unlabeled document using the source-only or co-training systems that had already been trained in the dialogue adaptation setup. Then, we sampled 100 documents with the *worst* confidence scores, because such data are unlikely to be selected in bootstrapping and accurately parsed. Finally, we fine-tuned each model on the 100 actively-labeled data. We also used random confidence (uncertainty) measures as the baseline, whose results are averaged over 5 trials. Table 5 shows that, even though only 100 dialogues were annotated manually, AL improved the performance significantly, which was difficult to achieve by bootstrapping alone. Annotating highly uncertainty data is more effective than annotating randomly sampled dialogues. We can also see that the combination of bootstrapping and AL achieves higher performance than the source-only model with AL, suggesting that bootstrapping and AL can be complementary and that bootstrapping is useful to identify potentially useful data in AL. The performance improvement could be further increased by repeating bootstrapping and AL alternatively, which is worth investigating in the future.

6.6 Summary and Recommendations

Here, we summarize what we have learned from the experiments and push the findings a step further in order to provide practical guidelines for out-of-domain discourse dependency parsing.

1. Bootstrapping improves out-of-domain discourse dependency parsing significantly and consistently in various adaptation scenarios. In particular, we recommend co-training with the arc-factored and shift-reduce models because co-training tends to be more effective and more efficient in training than tri-training variants.

2. A labeled source dataset that is as close as possible to the target domain is preferable to suppress the domain-shift problem. The labeled source dataset should also follow the the same annotation framework with the target domain (e.g., definitions of EDUs and discourse relation classes).

3. It is reasonable to use the models' predictive probability as the confidence measure to filter out noisy pseudo labels, because the confidence scores correlate with the accuracy of the pseudo labels. In particular, we recommend the Rank-above-k criterion because, unlike Rank-diff-k, k is independent of the number of unlabeled data. However, since the accurately predicted pseudo labels are biased towards simpler documents, the parsing accuracy on more complex documents is difficult to improve even with bootstrapping.

4. The low-coverage problem of pseudo labels is not alleviated by increasing the number of unlabeled target documents. We recommend manually annotating a small amount of target documents using active learning and combining it with bootstrapping.

7 Conclusion

In this paper, we investigated the effectiveness and limitation of bootstrapping methods in unsupervised domain adaptation of BERT-based discourse dependency parsers. The results demonstrate that bootstrapping is effective significantly and consistently in various adaptation scenarios. However, regardless of the tuned confidence measures and sample selection criteria, the bootstrapping methods have a difficulty in generating both diverse and accurate pseudo labels, which is

the current limiting factor in further improvement. This low-coverage problem cannot be mitigated by just increasing the unlabeled corpus size. We confirmed that the active learning can be the effective solution to this problem.

We have a limitation in this study: Our experiments use only English documents. Although bootstrapping and discourse parsing models are language-independent at the algorithmic level, experiments of domain adaptation require labeled datasets on both source and target domains for training and evaluation. In order to investigate the universality and language-dependence features of the bootstrapping methods in various languages, it is necessary to develop discourse treebanks in a variety of languages.

In the future, we will expand the COVID19-DTB dataset with additional biomedical abstracts to facilitate the exploration and application of discourse parsing technologies to biomedical knowledge acquisition.

Acknowledgments

We would like to thank the action editor and three anonymous reviewers for their thoughtful and insightful comments, which we found very helpful in improving the paper. This work was supported by JSPS KAKENHI 21K17815. This work was also partly supported by JST, AIP Trilateral AI Research, grant number JPMJCR20G9, Japan.

References

Steven Abney. 2002. Bootstrapping. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. https://doi.org/10.3115/1073083.1073143

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 928–937. https://doi.org/10.18653/v1/D15-1109

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Afantenos Stergos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 2721–2727.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019a. Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 640–645. https://doi.org/10.18653/v1/P19-1061

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019b. Weak supervision for learning discourse structure. 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 2296–2305. https://doi.org/10.18653/v1/D19-1234

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620. https://doi.org/10.18653/v1/D19-1371

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218. https://doi.org/10.18653/v1/D15-1263

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pages 92–100. https://doi.org/10.1145/279943.279962

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *Technical Report ISI-TR-545*. University California Information Sciences Institute.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2001. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In *Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*. https://doi.org/10.3115/1118078.1118083

Danqi Chen and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750. https://doi.org/10.3115/v1/D14-1082

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 (NAACL-HLT)*, pages 4171–4186.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL2016)*, pages 1998–2008. https://doi.org/10.18653/v1/P16-1188

Vanessa Wei Feng and Graema Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 511–521.

Elisa Ferracane, Su Wang, and Raymond J. Mooney. 2017. Leveraging discourse information effectively for authorship attribution. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 584–593.

Grigorii Guz and Giuseppe Carenini. 2020. Coreference for discourse parsing: A neural approach. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, pages 160–167.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010a. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 399–409.

Hugo Hernault, Helmut Prendinger, David a. DuVerle, and Mitsuru Ishizuka. 2010b. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, 1(3):1–33. https://doi.org/10.5087/dad.2010.003

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference of Empirical Methods in Natural Language Processing (EMNLP)*, pages 1515–1520.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5010–5016.

Zhongqiang Huang and Mary Harper. 2009. Self-training PCFG grammars with latent annotations across languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 832–841. https://doi.org/10.3115/1699571.1699621

Patrick Huber and Giuseppe Carenini. 2019. Predicting discourse structure using distant supervision from sentiment. 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 2306–2316. https://doi.org/10.18653/v1/D19-1235

Ofir Israeli, Adi Beth-Din, Nir Paran, Dana Stein, Shirley Lazar, Shay Weiss, Elad Milrot, Yafit Atiya-Nasagi, Shmuel Yitzhaki, Orly Laskar, and Ofir Schuster. 2020. Evaluating

the efficacy of RT-qPCR SARS-CoV-2 direct approaches in comparison to RNA extraction. *BioRxiv preprint 2020.06.10.144196v1*. https://doi.org/10.1101/2020.06.10.144196

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 977–986. https://doi.org/10.3115/v1/P14-1092

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13–24.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 996–1005.

Kailang Jiang, Giuseppe Carenini, and Raymond T. Ng. 2016. Training data enrichment for infrequent discourse relations. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pages 2603–2614.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77. https://doi.org/10.1162/tacl_a_00300

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435. https://doi.org/10.1162/COLI_a_00226

Shafiq Joty, Giuseppe Carenini, Raymond T. Ng, and Yashar Mehdad. 2013. Combining intra- and multi-sentential rhetorical parsing for document-level discourse analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 486–496.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5851–5861. https://doi.org/10.18653/v1/P19-1586

Eliyahu Kiperwasser and Yoav Goldberg. 2016. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327. https://doi.org/10.1162/tacl_a_00101

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down RST parsing utilizing granularity levels in documents. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, pages 8099–8106. https://doi.org/10.1609/aaai.v34i05.6321

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2021. Improving neural RST parsing model with silver agreement subtrees. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1600–1612. https://doi.org/10.18653/v1/2021.naacl-main.127

Naoki Kobayashi, Tsutomu Hirao, Kengo Nakamura, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2019. Split of merge: Which is better for unsupervised RST parsing? In *Proceedings of the 2019 Conference of Empirical Methods in Natural Language Processing (EMNLP)*, pages 5797–5802. https://doi.org/10.18653/v1/D19-1587

Fajri Koto, Jey Han Lan, and Timothy Baldwin. 2021. Top-down discourse parsing via sequence labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 715–726. https://doi.org/10.18653/v1/2021.eacl-main.60

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and

Bing Qin. 2020. Molweni: A challenge multiparty dialogue-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 2642–2652.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014a. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference of Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069.

Qi Li, Tianshi Li, and Baobao Chang. 2016a. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference of Empirical Methods in Natural Language Processing (EMNLP)*, pages 362–371. https://doi.org/10.18653/v1/D16-1035

Sujian Li, Liang Wang, Ziqiang Cao, and Wenjie Li. 2014b. Text-level discourse dependency parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 25–35.

Zhenghua Li, Min Zhang, Yue Zhang, Zhanyi Liu, Wenliang Chen, Hua Wu, and Heifeng Wang. 2016b. Active learning for dependency parsing with partial annotation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 344–354.

Yang Liu and Mirella Lapata. 2018. Learning structured text representations. *Transactions of the Assocication for Computational Linguistics*, 6:63–75. https://doi.org/10.1162/tacl_a_00005

Annie Louis, Aravind Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summarization. In *Proceedings of the SIGDIAL 2010 Conference*, pages 147–156.

Ryan Lowe, Nissan Pow, Lulian Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIDDIAL)*, pages 285–294.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281. https://doi.org/10.1515/text.1.1988.8.3.243

Daniel Marcu. 1999. A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 365–372.

David McClosky, Eugene Charniak, and Mark Johnson. 2006. Effective self-training for parsing. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 152–159.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (EMNLP-HLT)*, pages 523–530. https://doi.org/10.3115/1220575.1220641

Mathieu Morey, Philippe Muller, and Nicholas Asher. 2018. A dependency perspective on RST discourse parsing and evaluation. *Computational Linguistics*, 44(2):197–235. https://doi.org/10.1162/coli_a_00314

Philippe Muller, Stergos Afantenos, Pascal Denis, and Nicholas Asher. 2012. Constrained decoding for text-level discourse parsing. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1883–1900.

Noriki Nishida and Hideki Nakayama. 2020. Unsupervised discourse constituency parsing using Viterbi EM. *Transactions of the Association for Computational Linguistics*, 8:215–230. https://doi.org/10.1162/tacl_a_00312

Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the ACL Workshop Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57. https://doi.org/10.3115/1613148.1613156

Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 99–109. https://doi.org/10.18653/v1/N16-1013

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 1171–1182. https://doi.org/10.18653/v1/E17-1110

Alexander J. Ratner, Cristopher M. De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pages 3574–3582.

Roi Reichart and Ari Rappoport. 2007. Self-training for enhancement and domain adaptationa of statistical parsers train on small datasets. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 616–623.

Sebastian Ruder and Barbara Plank. 2018. Strong baselines for neural semi-supervised learning under domain shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1044–1054. https://doi.org/10.18653/v1/P18-1096

Kenji Sagae. 2009. Analysis of discourse structure with syntactic dependencies and data-driven shift-reduce parsing. In *Proceedings of the 11th International Workshop on Parsing Technology (IWPT)*, pages 81–84. https://doi.org/10.3115/1697236.1697253

Kuniaki Saito, Toshitaka Ushiku, and Tatsuya Harada. 2017. Asymmetric tri-training for unsupervised domain adaptation. In *Proceedings of The 34th International Conference on Machine Learning (ICML)*, pages 2988–2997.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 7007–7014. https://doi.org/10.1609/aaai.v33i01.33017007

Anders Søaard and Christian Rishøj. 2010. Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1065–1073.

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003a. Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 236–243. https://doi.org/10.3115/1073445.1073476

Mark Steedman, Miles Osborne, Anoop Sarkar, Stephen Clark, Rebecca Hwa, Julia Hockenmaier, Paula Ruhlen, Steven Baker, and Jeremiah Crim. 2003b. Bootstrapping statistical parsers from small datasets. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 331–338. https://doi.org/10.3115/1067807.1067851

Jun Suzuki and Hideki Isozaki. 2008. Semi-supervised sequential labeling and segmentation using giga-word scale unlabeled data. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 665–673.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 735–736. https://doi.org/10.1145/1277741.1277883

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Douglas Burdick, Darrin Eide, Kathryn Funk,

Yannis Katsis, Rodney Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. *arXiv preprint arXiv: 2004.10706v4*.

Yizhong Wang, Sujian Li, and Honfeng Wang. 2017. A two-stage parsing method for text-level discourse analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 184–188. https://doi.org/10.18653/v1/P17-2029

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D18-1116

David Weiss, Chris Alberti, Michael Collins, and Slav Petrov. 2015. Structured training for neural network transition-based parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 323–333. https://doi.org/10.3115/v1/P15-1032

Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-aware neural extractive text summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5021–5031.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency treebank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 444–449. https://doi.org/10.18653/v1/P18-2071

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 189–196. https://doi.org/10.3115/981658.981684

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1834–1839. https://doi.org/10.3115/v1/D14-1196

Liwen Zhang, Ge Wang, Wenjuan Han, and Kewei Tu. 2021. Adapting unsupervised syntactic parsing methodology for discourse dependency parsing. 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 5782–5794. https://doi.org/10.18653/v1/2021.acl-long.449

Longyin Zhang, Yuqing Xing, Fang Kong, Peifeng Li, and Guodong Zhou. 2020. A top-down neural architecture towards text-level parsing of discourse rhetorical structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6386–6395. https://doi.org/10.18653/v1/2020.acl-main.569

Yan Zhou and Sally Goldman. 2004. Democratic co-learning. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 17(11):1529–1541. https://doi.org/10.1109/TKDE.2005.186