Efficient Argument Structure Extraction with Transfer Learning and Active Learning

Xinyu Hua
Bloomberg
New York, NY
xhua22@bloomberg.net

Lu Wang
Computer Science and Engineering
University of Michigan
Ann Arbor, MI
wangluxy@umich.edu

Abstract

The automation of extracting argument structures faces a pair of challenges on (1) encoding long-term contexts to facilitate comprehensive understanding, and (2) improving data efficiency since constructing high-quality argument structures is time-consuming. In this work, we propose a novel context-aware Transformer-based argument structure prediction model which, on five different domains, significantly outperforms models that rely on features or only encode limited contexts. To tackle the difficulty of data annotation, we examine two complementary methods: (i) *transfer learning* to leverage existing annotated data to boost model performance in a new target domain, and (ii) *active learning* to strategically identify a small amount of samples for annotation. We further propose model-independent sample acquisition strategies, which can be generalized to diverse domains. With extensive experiments, we show that our simple-yet-effective acquisition strategies yield competitive results against three strong comparisons. Combined with transfer learning, substantial F1 score boost (5-25) can be further achieved during the early iterations of active learning across domains.

1 Introduction

Identifying and understanding the argumentative discourse structure in text has been a critical task in argument mining (Peldszus and Stede, 2013; Cabrio and Villata, 2018; Lawrence and Reed, 2019; Li et al., 2020). It plays an important role of discovering the central theses and reasoning process across a wide spectrum of domains, from formal text such as legal documents (Palau and Moens, 2009; Lippi and Torroni, 2016; Poudyal et al., 2020) and scientific literature (Mayer et al., 2020; Fergadis et al., 2021; Al Khatib et al., 2021), to online posts and discussions (Cardie et al., 2008; Boltužić and Šnajder, 2014; Park and Cardie, 2014; Habernal and Gurevych, 2017; Hua and Wang, 2017). Here we focus on automatic **argumentative relation prediction**—given any proposition in a document, predict the existence and polarity (support or attack) of relation from any other proposition within a specified context window. One major challenge resides in capturing *long-term dependencies*. As illustrated in Fig. 1, propositions with an argumentative relation are often separated by a large text span, requiring the understanding of a longer context (Nguyen and Litman, 2016; Opitz and Frank, 2019).

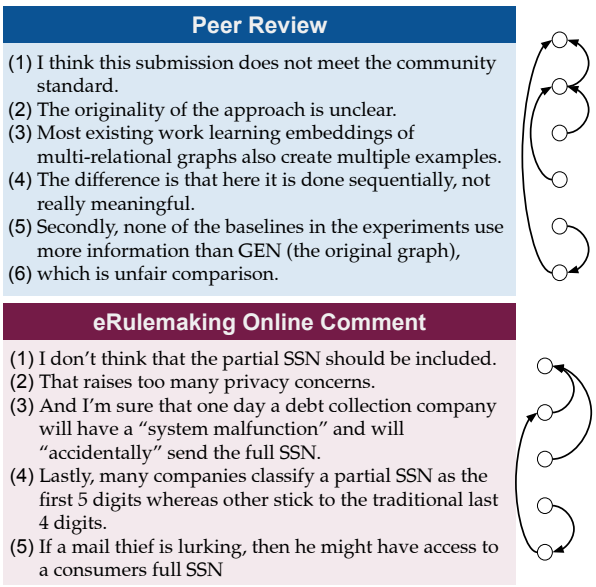

Figure 1: Excerpts of arguments in peer reviews and online comments. On the right, argumentative structure is labeled as support relations among propositions. Despite differences in topics and vocabularies, we see similar structural patterns with long-term dependencies, motivating learning transferable representations across domains.

Existing methods for this important task are often time-consuming, as they require at least three steps (Nguyen and Litman, 2016; Stab and Gurevych, 2017; Niculae et al., 2017; Mayer et al., 2020): (1) acquiring high-quality labels from do-

Findings of the Association for Computational Linguistics: ACL 2022, pages 423 - 437

main experts, (2) manually designing customized features to address long dependencies and encode task-specific language, and (3) model training. To exacerbate the challenge, the resulting models are hardly generalizable to new domains.

Consequently, our main goal is to design an *easy-to-use* framework that can facilitate researchers and practitioners to build argument structure extraction models for *new domains rapidly* and *accurately*. To this end, we first propose a novel *context-aware* argument relation prediction model, which can be directly fine-tuned from pre-trained Transformers (Vaswani et al., 2017; Liu et al., 2019). For a given proposition, the model encodes a broad context of neighboring propositions in the same document, and predicts whether each of them supports, attacks, or has no relation to the original one. By contrast, prior work only encodes pairwise propositions while ignoring contexts (Mayer et al., 2020).

Moreover, while training on a large labeled corpus has become the *de facto* method for neural models, labeling argument structures is a laborious process even for experienced annotators with domain knowledge (Green, 2014; Saint-Dizier, 2018; Lippi and Torroni, 2016). Our second goal is to investigate *efficient model training*, by using fewer samples for a new domain. We study the following two complementary techniques: (i) **Transfer learning** (TL) adapts models trained on existing annotated data in a different domain, or leverages unlabeled in-domain data for better representation learning. (ii) **Active learning** (AL) strategically selects *samples in the new domain* based on a sample acquisition strategy with the goal of optimizing training performance. This process is often done in multiple rounds within a given budget (Settles, 2009). As pointed out by Lowell et al. (2019), model-specific selection methods may not generalize across successor models and domains. We thus design model-independent strategies to encourage the inclusion of *unseen words*, and sentences with *discourse markers*. Both are easy to implement and incur little computation cost. We compare them with popular methods based on uncertainty (Lewis and Gale, 1994; Houlsby et al., 2011) and sample diversity (Sener and Savarese, 2018).

For experiments, we release **AMPERE++**[1], the first dataset in the peer review domain labeled with argument relations. Our annotation process involves over 10 months of training and multi-round sessions with experienced annotators, finally yielding 3,636 relations over 400 reviews originally collected in our prior work (Hua et al., 2019). It has the highest overall relation density and the most attack relations, compared to prior datasets (Table 1). We also evaluate on four other datasets covering diverse topics, including Essays (Stab and Gurevych, 2017), AbstRCT (Mayer et al., 2020) for biomedical paper abstracts, ECHR (Poudyal et al., 2020) for case-law documents, and the Cornell eRulemaking Corpus (CDCP) (Park and Cardie, 2018) for online comments on public policies. Our second data contribution comprises three large collections of unlabeled samples tailored for self-supervised pretraining for the first three domains.

Drawing from extensive experiment results, we make the following observations: (1) Our proposed model, which can encode longer contexts, yields better argument relation prediction results than comparisons or variants that operate over limited contexts (§6.1). (2) TL substantially improves performance for target domains when less labeled data is available. For example, for ECHR and CDCP, using AMPERE++ as the source domain, with only half of the target domain training data, the model achieves better F1 scores than non-transferred model trained over the entire training set (§6.2). This also highlights the value of our AMPERE++ data. (3) Among AL methods, our newly proposed model-independent acquisition strategies yield competitive results against comparisons that require significantly more computations (§6.3). (4) TL further improves all AL setups and narrows the gaps among strategies (§6.3).

2 Related Work

Argument Structure Extraction. Analyzing argumentation in natural language text has seen rapid growth (Lippi and Torroni, 2016; Cabrio and Villata, 2018; Lawrence and Reed, 2019), yet the most challenging aspect of it is to extract the structures among diverse argument components. Conceptually, the structure extraction model needs to address two subtasks: (1) determining which propositions are targeted (head detection), and (2) identifying the argumentative relations towards the head propositions. Early work (Peldszus and Stede, 2013, 2015) takes inspiration from discourse parsing. While practically argument relations can be dis-

[1]Data and code are available at https://xinyuhua.github.io/Resources/acl22/.

persed across the text, contrary to assumptions in common discourse theory (Mann and Thompson, 1988; Webber et al., 2019). More recent work considers all pairwise combinations of propositions (Stab and Gurevych, 2014; Niculae et al., 2017; Mayer et al., 2020), which incurs expensive computations for long documents. Our model encodes a sequence of propositions and extract their labels in one forward pass, leading to much reduced training and inference complexity while allowing access to more contexts.

Transfer Learning for Structured Prediction. Collecting human annotations for structured tasks is costly, especially when discourse-level understanding and domain expertise are required (Mieskes and Stiegelmayr, 2018; Schulz et al., 2019; Poudyal et al., 2020). It is thus desirable to reuse existing labels from a similar task, and transfer learning (TL) is often employed. It can be divided into two broad categories (Pan and Yang, 2009): (1) *Transductive* approaches adapt models learned from a labeled source domain to a different target domain over the same task, and have shown promising results for discourse (Kishimoto et al., 2020) and argument (Chakrabarty et al., 2019; Accuosto and Saggion, 2019) related tasks. (2) *Inductive* methods aim to leverage unlabeled data, usually in the same domain as the target domain, and have gained popularity with the pre-training and fine-tuning paradigm using Transformer models (Devlin et al., 2019; Gururangan et al., 2020). We study both types in this work, with a particular focus on transductive approaches where the effect of different source domains are compared.

Active Learning (AL) has been explored in many NLP problems including named entity recognition (Tomanek and Hahn, 2009; Shen et al., 2018), text classification (Tong and Koller, 2001; Hoi et al., 2006), and semantic parsing (Iyer et al., 2017; Duong et al., 2018). Unlike the traditional supervised setting where training data is sampled beforehand, AL allows the learning system to actively select samples to maximize the performance, subject to an annotation budget (Settles, 2009; Aggarwal et al., 2014). Common AL strategies are either based on model uncertainty (Houlsby et al., 2011; Yuan et al., 2020), or promoting the diversity in sample distribution (Bodó et al., 2011; Sener and Savarese, 2018). However, both paradigms require coupling sampled data with a specific learned model, which may cause subpar performance by a successor model (Lowell et al., 2019). We propose model-independent acquisition strategies that are faster to train and do not rely on any model.

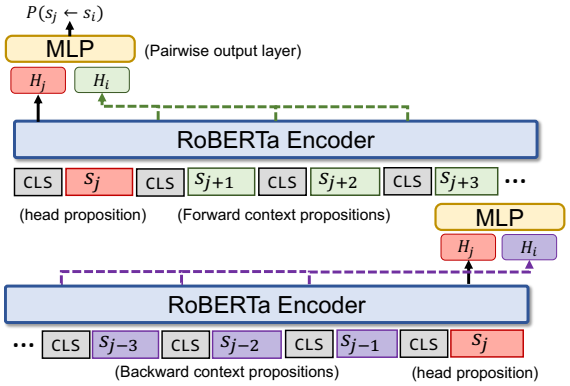

Figure 2: Our context-aware argument relation prediction model. For each head proposition $\boldsymbol{s}_j$, we encode both the backward (purple) and forward (green) contexts. H_j, the last layer states, represents proposition $\boldsymbol{s}_j$. H_i, where i can be $j \pm 1, j \pm 2, \ldots, j \pm L$ (L is the window size), is concatenated with H_j and fed into the pairwise output layer, to yield the probability of $\boldsymbol{s}_j \leftarrow \boldsymbol{s}_i$.

3 Argument Relation Prediction Model

Task Formulation. Given a document that is segmented into a list of propositions, our task is to predict the existence of a `support` or `attack` link $\boldsymbol{s}_j \leftarrow \boldsymbol{s}_i$ between propositions $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. Here the targeted proposition $\boldsymbol{s}_j$ is the *head*, and $\boldsymbol{s}_i$ is the *tail*. Our end-to-end model considers all proposition pairs. We also consider a simplified setting, where head propositions are given a priori.

A Context-aware Model. Fig. 2 depicts our model: It is built on top of the RoBERTa encoder (Liu et al., 2019) which reads in a sequence of tokens. It contains stacked layers with bidirectional multi-headed self-attentions. Different from prior work that only encodes single propositions, given a head proposition $\boldsymbol{s}_j$, we concatenate it with its surrounding context, including the L propositions before and after it. Propositions are separated by `[CLS]` tokens. We use their last layer's states, denoted as H_j, to represent $\boldsymbol{s}_j$. Other propositions within the window defined by L then become candidates for tail propositions.

After encoding, each tail candidate representation H_i is concatenated with the head representation H_j to form the input to the output layer, with the final prediction formulated as:

$$P(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i) = \text{softmax}(\tanh([H_j; H_i] \cdot \boldsymbol{W}_1) \cdot \boldsymbol{W}_2) \quad (1)$$

where y_r corresponds to three classes: `support`, `attack`, and `no-rel` if there is no link. $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are trainable parameters. Dropout (Srivastava et al., 2014) is added between layers.

Training objective is cross-entropy loss over the labels of pairwise propositions within the context window. Our simplified setting reduces the prediction complexity from $\mathcal{O}(n^2)$ (Mayer et al., 2020) to $\mathcal{O}(nL)$, with n being the proposition count.

4 Active Learning Strategies

One major goal of this work is to explore AL solutions that can reduce the amount of samples for annotation, since labeling such a dataset can be the most laborious part of argument structure understanding. We consider a pool-based AL scenario (Settles, 2009), where labels for the training set $\mathcal{U}$ are assumed to be unavailable initially. The learning procedure is carried out in T iterations. In the t-th iteration, b samples are selected using a given acquisition strategy. These samples are labeled and added into the labeled pool to comprise $\mathcal{D}_t$, on which a model $\mathcal{M}_t$ is then trained.

4.1 Comparison Methods

For baselines, we consider **RANDOM-PROP**, which samples b propositions from the unlabeled training set with uniform distribution. Its variant, **RANDOM-CTX**, instead samples at the context level — i.e., for a given head, its entire forward or backward context of L propositions are sampled as a whole, until the total number of propositions reaches b.

The **MAX-ENTROPY** (Lewis and Gale, 1994; Joshi et al., 2009) method selects the most uncertain samples, based on the entropy score $\mathcal{H}(\cdot)$ using the model trained in the previous iteration:

$$\mathcal{H}(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i) = -\sum_r P(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i) \log P(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i) \quad (2)$$

where $P(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i)$ is the predicted probability of a relation label (Eq. 1).

Bayesian Active Learning by Disagreement (**BALD**) (Houlsby et al., 2011) is another common approach to exploit the uncertainty of unlabeled data by applying dropout at test time for multiple runs over the same sample, and picks ones with higher disagreement:

$$\arg\max_{\boldsymbol{s}_i} \mathcal{H}(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i) - \mathbb{E}_\theta[\mathcal{H}(y_r|\boldsymbol{s}_j, \boldsymbol{s}_i, \theta)] \quad (3)$$

Uncertainty-based methods are at risk of selecting "outliers" or alike samples (Settles, 2009). To encourage diversity of the selected samples, we consider **CORESET** (Sener and Savarese, 2018), which enlarges differences among samples and achieves competitive performance in many vision tasks. At a high level, each sample is represented as a vector, e.g., we use the proposition representation H_i. A random set of b samples are selected for labeling in the first iteration. In each subsequent iteration t, data points in the labeled pool $\mathcal{D}_{t-1}$ are treated as cluster centers, and the sample with the greatest L_2 distance from its nearest cluster center is selected. This process is repeated b times to build the new labeled pool $\mathcal{D}_t$.

4.2 Model-independent Acquisition Methods

One risk in AL is that samples selected by a model might not be useful for future models (Lowell et al., 2019). This motivates our design of *model-independent* acquisition methods. Our first method, **NOVEL-VOCAB**, promotes propositions with more unseen words. Assuming the frequency of a word w in the labeled pool is $\mathcal{V}(w)$, the novelty score for an unlabeled sample $\boldsymbol{s}_i$ is computed as:

$$\text{novelty-score}(\boldsymbol{s}_i) = \sum_{w_t \in \boldsymbol{s}_i} \frac{f_{i,t}}{(1 + \mathcal{V}(w_t))} \quad (4)$$

where $f_{i,t}$ is the frequency of word w_t in sample $\boldsymbol{s}_i$. Samples with the highest novelty scores are selected for labeling. If a proposition has a high word overlap with samples in the labeled pool, the denominator $\mathcal{V}(w_t)$ will be high, and this sample is less likely to be chosen.

Our second method, **DISC-MARKER**, aims to select more relation links by matching any of the following 18 prominent discourse markers from PDTB manual (Webber et al., 2019) (matching statistics are in Appendix A.1).[2] For comparison, we also show a complementary approach **NO-DISC-MARKER**, which samples propositions *without* any of those discourse markers.

because	therefore	however
although	though	nevertheless
nonetheless	thus	hence
consequently	for this reason	due to
in particular	particularly	specifically
in fact	actually	but

[2]When matched sentences exceed selection budget, we randomly sample with equal probabilities.

	AMPERE++	Essays	AbstRCT	ECHR	CDCP
# Doc.	400	402	700	42	731
# Tok.	190k	147k	236k	177k	89k
# Prop.	10,386	12,373	5,693	6,331	4,932
# Supp.	3,370	3,613	2,402	1,946	1,426
# Att.	266	219	70	0	0
# Head	2,268	1,707	1,138	741	1,037
Density	21.8%	13.8%	20.0%	11.7%	21.0%

Table 1: Statistics of five datasets, including our AMPERE++ data with newly annotated relations on AMPERE (Hua et al., 2019). We report the total numbers of documents (# Doc.), tokens (# Tok.), propositions (# Prop.), support (# Supp.) and attack (# Att.) relations, unique head propositions (# Head), and relation density as the percentage of propositions that are supported or attacked by at least one proposition.

5 Datasets and Domains

We experiment with five datasets from distinct domains, with key statistics listed in Table 1. Below we outline data collection and annotation, notable preprocessing steps, and data splits.

Domain 1: Peer Reviews (New Annotation). We first annotate argument relations on AMPERE (Hua et al., 2019), which consists of 400 ICLR 2018 paper reviews collected from OpenReview. Each review has been annotated with segmented propositions and corresponding types (i.e., *evaluation*, *request*, *fact*, *reference*, and *quote*). We augment this dataset by labeling the support and attack relations among the propositions. This new dataset is called **AMPERE++**.

We hire three proficient English speakers to annotate the entire dataset in multiple rounds. During annotation, they are displayed with the propositions along with their types. We impose two constraints. (1) Each proposition can only support or attack at most one other proposition. (2) Factual propositions (*fact*, *reference*, *quote*) cannot be supported or attacked by subjective ones (*evaluation*, *request*). Similar rules are used by Park and Cardie (2018). We include detailed guidelines in Appendix B. For quality control and disagreement resolution, the annotators are joined by a fourth judge after each round, where they discuss samples with different labels to reach agreement.

The resulting dataset contains 3,636 relations from 400 reviews with a substantial inter-annotator agreement score of 0.654 (Fleiss' κ). Following our prior work (Hua et al., 2019), we use 300 reviews for training, 20 for validation, and 80 for test. We also collect 42k reviews from OpenReview for ICLR 2019-2021, UAI 2018, and NeurIPS 2013-2020, which are used in the self-supervised learning experiments for improving representation learning.

Domain 2: Essays. Our second dataset is based on the essays curated by Stab and Gurevych (2017) from `essaysforum.com`. Argumentative propositions are identified at the sub-sentence level and labeled as "*premise*", "*claim*", or "*major claim*". Support and attack relations are annotated from a premise to a claim or to another premise. The link cannot cross paragraph boundaries, highlighting the dataset's focus on relations close by.

We split the original training set into 282 essays for training and 40 for validation. The remaining 80 are reserved for test. Similarly, we also download 26K essays from the same online forum for self-supervised representation learning.

Domain 3: Biomedical Paper Abstracts. Next, we use the **AbstRCT** corpus (Mayer et al., 2020), which contains 700 paper abstracts retrieved from PubMed.[3] The primary subjects are Randomized Controlled Trials of diseases. Notably, AbstRCT has much fewer propositions and relations than the previous two datasets, due to the factual nature of paper abstracts.

Following Mayer et al. (2020), we use 350 abstracts for training, 50 for validation, and 300 for test. We employ the 133K unlabeled abstracts released by Cohan et al. (2018) for self-supervision.

Domain 4: Legal Documents. Legal texts are studied in the early work of argument mining (Palau and Moens, 2009; Lippi and Torroni, 2016). We choose the **ECHR** corpus (Poudyal et al., 2020), containing 42 recently-annotated case-law documents of the European Court of Human Rights. The authors define an argument structure as a list of premises and a conclusion. We consider each premise as linked to the corresponding conclusion. The dataset is split into 27 documents for training, 7 for validation, and 8 for test.

Domain 5: Online User Comments. Finally, we include the Cornell eRulemaking Corpus (Park and Cardie, 2018), extracted from an online forum where the public argues for or against proposed rules. The 731 annotated comments are mostly related to the Consumer Debt Collection Practices

[3] `https://pubmed.ncbi.nlm.nih.gov/`

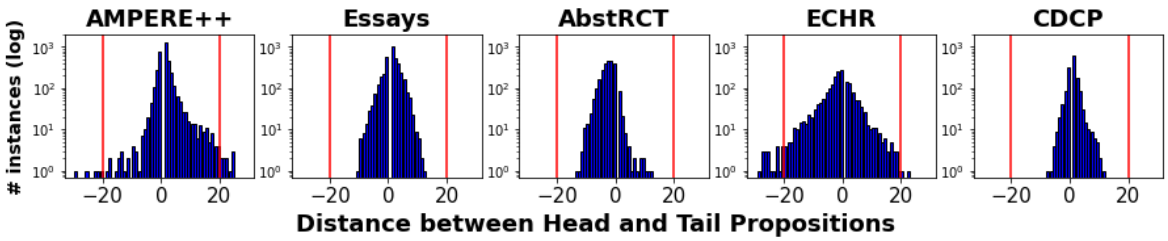

Figure 3: Distribution of distance (measured by number of propositions) between head-tail pairs across five domains. Positive values indicate that the tail appears after the head in the document, and vice versa.

rule (**CDCP**), and is annotated with support relations only. We adopt the original splits: 501 for training, 80 for validation, and 150 for test. On average, there are less than two relation links per comment, and only 21% of the propositions are supported.

Head-tail Distance Distribution. Recall that our context-aware model only encodes context propositions up to a fixed window size. Although this setup neglects some relation links, we show in Fig. 3 that a large enough window size (e.g., 20) is sufficient to cover all (Essays, CDCP, AbstRCT) or over 98% (AMPERE++, ECHR) of all relations.

Fig. 3 further highlights domain-specific patterns. AMPERE++ and CDCP are skewed to the right, indicating reviewers and online users tend to put their claims upfront with supporting arguments appearing later. On the contrary, paper abstracts (AbstRCT) usually describe premises first and then draw conclusions. Essays and ECHR have more balanced distributions between both directions.

Proposition Length and Label Distribution. Due to differences in argument schemes, proposition length varies considerably across domains. AbstRCT has the longest propositions with an average of 45 tokens. Consequently, the actual encoder input may contain less than 20 propositions due to the maximum token limit. Under our context-aware encoding, the ratio of positive samples (`support` or `attack`) is boosted to 29% because they are less likely to be truncated due to the relative proximity to head propositions (Fig. 3). The other four domains have similar positive ratios, ranging from 6% (AMPERE++) to 17% (CDCP).

Existing relation prediction methods (Stab and Gurevych, 2017; Niculae et al., 2017; Mayer et al., 2020) label all pairwise propositions within the same document, leading to much lower positive ratios, especially for ECHR where documents are long. In §6.1 we show that such unbalanced distribution poses difficulties for traditional methods.

6 Experiments and Results

In this section, we design experiments to answer the following questions. (1) To which degree is the context-aware model better at identifying argumentative relations (§6.1)? (2) How much improvement can transfer learning (TL) make when different source domains are considered for a target domain (§6.2)? (3) Does unlabeled in-domain data help downstream tasks using self-supervised pretraining and inductive transfer learning (§6.2)? (4) How do active learning (AL) strategies perform on relation prediction and whether combining transfer learning leads to further performance boost (§6.3)?

Evaluation is based on macro-F1 scores as done in prior work (Stab and Gurevych, 2017; Niculae et al., 2017). For tasks without attack labels (ECHR and CDCP), the macro average is calculated over `support` and `no-rel` only, otherwise it is averaged over three classes. Each setup is run five times with different random seeds, and the average scores on test sets are reported.

Implementation of our models is based on the Transformer (Wolf et al., 2020). Our encoder is RoBERTa-base (Liu et al., 2019), which has 12 layers with a hidden size of 768. We apply dropout (Srivastava et al., 2014) with a probability of 0.1 for the output MLP layer. We use the Adam optimizer (Kingma and Ba, 2015) with 16 sequences per batch. We hypertune our proposed argument relation prediction model with different number of maximum training epochs $\{5, 10, 15\}$, warmup steps $\{0, 1000, 5000\}$, learning rate $\{$1e-5, 1e-6, 5e-5$\}$, and scheduler $\{$`constant`, `linear`$\}$. The best validation result is achieved with 15 epochs, 5000 warmup steps, 1e-5 as learning rate, and the `constant` scheduler. We use this configuration for all model training experiments.

	AMPERE++	Essays	AbstRCT	ECHR	CDCP
SVM-linear	24.82	28.69	33.60	21.18	29.01
SVM-RBF	26.38	31.68	32.65	21.36	30.34
SEQPAIR	23.40	38.37	**66.96**	13.76	35.23
BENCHMARK	-	**73.30**	-	-	26.70
OURS (head given)					
$L = 5$	66.34	65.61	55.48	60.92	64.82
$L = 10$	75.69	69.41	59.27	67.51	69.47
$L = 20$	**77.64**	71.30	63.62	**70.82**	**70.37**
OURS (end-to-end)					
$L = 20$	74.34	67.68	63.73	61.35	63.13

Table 2: F1 scores for argument relation prediction. Each entry is averaged over five runs with different random seeds. The best result for each dataset is **bolded**. Our context-aware model outperforms both baselines except for AbstRCT. The difference between *head given* and *end-to-end* is close, suggesting that the key challenge for structure extraction lies in relation prediction. Our model performance improves when larger window size L is used.

6.1 Supervised Learning Results

We first evaluate our model with the standard supervised learning over the full training set using varying window sizes. We assume the heads are given at both training and inference, except for the end-to-end setting.

Comparisons. We implement an **SVM** with features adapted from Table 10 of Stab and Gurevych (2017), except for features specific to the essays domain (e.g., whether a proposition is in the introduction). We experiment with both linear and radial-basis function (RBF) kernels, with regularization coefficients tuned on validation. More details can be found in Appendix A.2.

SEQPAIR is based on the sequence pair classification setup (Devlin et al., 2019) using the pretrained RoBERTa. Each pair of head and tail is concatenated and segmented with the `[SEP]` token. The `[CLS]` token is prepended to the beginning of the sequence and used for classification. This setup resembles the model in Mayer et al. (2020).

We further compare with two dataset-specific **BENCHMARK** models: Stab and Gurevych (2017) use a rich set of features tailored for essays to train SVMs, and Niculae et al. (2017) employ structured SVMs on CDCP.

Results. As shown in Table 2, our context-aware model outperforms the comparisons except for Essays and AbstRCT. The feature-rich SVM marginally outperforms our model, though the features are not generalizable to new domains. As mentioned in §5, AbstRCT has much higher positive ratio than other domains. This indicates that our model is more robust against unbalanced training data than the pairwise approach.

The performance drop for end-to-end models are marginal in most cases, underscoring relation prediction as the key challenge for structure extraction, which the simplified setup has to tackle as well.

6.2 Transfer Learning Results

Results in the previous section show large performance discrepancies among different domains. For instance, domains with few labeled samples, such as AbstRCT and CDCP, lead to worse performance. Moreover, annotating argument structures for some domains is even more involved, e.g., Poudyal et al. (2020) hired three lawyers to annotate ECHR legal documents. We hypothesize that basic reasoning skills for understanding argument structures can be shared across domains, thus we study transfer learning, a well-suited technique that leverages existing data with similar task labels (*transductive*) or unlabeled data of the same target domain (*inductive*). Concretely, we present thorough experiments of TL over all transfer pairs, where the model is first trained on the source domain and fine-tuned on the target domain.

Transductive TL. The upper half of Table 3 shows that *three out of four models transferred from AMPERE++ achieve better performance than their supervised learning counterparts in Table 2*. In particular, we observe more than 5 F1 points gains on ECHR and CDCP, which contain the least amount of labeled samples. However, when transferred from the four other datasets, performance occasionally drops. This can be due to the distinct language style and argumentative structure (AbstRCT), the source domain size (CDCP, ECHR), or the model's failure to learn good representations due to over-reliance on discourse markers (Essays). Overall, *AMPERE++ consistently benefits diverse domains for argument structure understanding, demonstrating its usage for future research*.

Inductive TL. Motivated by recent findings (Beltagy et al., 2019; Lee et al., 2020; Gururangan et al., 2020) that self-supervised pre-training over specific domains significantly improves downstream tasks, we also consider the *inductive* transfer learning setup with the following two objectives: (1) masked language

	AMPERE++	Essays	AbstRCT	ECHR	CDCP
SRC → TGT (Transductive TL)					
AMPERE++	–	73.84	63.42	**76.50**	**75.93**
Essays	77.93	–	60.62	68.72	74.11
AbstRCT	76.29	71.17	–	73.31	69.17
ECHR	77.69	70.82	47.91	–	69.30
CDCP	77.87	68.37	62.38	72.03	–
TGT-pret → TGT (Inductive TL)					
MLM	78.10	74.21	**64.48**	–	–
Context-Pert	79.01	68.36	59.47	–	–
SRC-pret → SRC → TGT					
AMPERE++	–	70.42	61.84	70.96	74.82
Essays	44.40	–	58.59	73.58	71.84
AbstRCT	76.25	69.26	–	70.93	71.67
TGT-pret → SRC → TGT					
AMPERE++	–	**74.90**	62.34	–	–
Essays	76.69	–	62.38	–	–
AbstRCT	**79.52**	73.09	–	–	–

Table 3: Results for transfer learning. First column denotes the source domain, the rest are target domains. The best result per column is in **bold**. Transfer learning that outperforms the in-domain training setup (Table 2, second last row) is highlighted in green. Notably, using AMPERE++ as the source domain yields better performance than the standard supervised setting. Overall, self-supervised pre-training can further benefit transductive transfer learning.

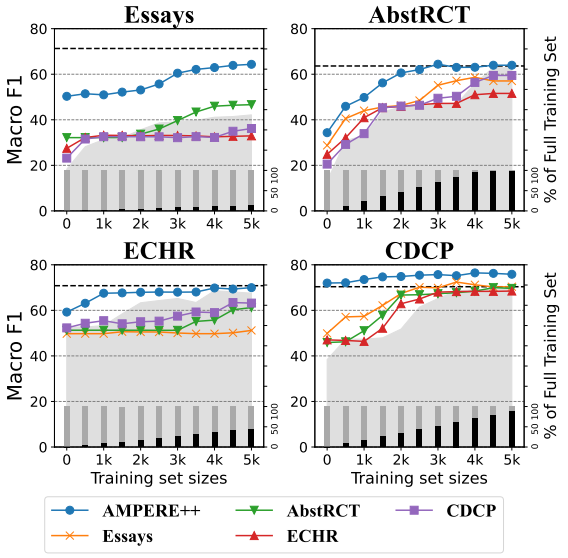

Figure 4: Macro F1 scores with limited training data. We sample training set from 0 to 5,000 samples, in an increment of 500. Bottom bars indicate the percentage of such subsets over the full training set. Scatter plots represent the transfer learning results from different source domains, with those from non-TL settings marked as shaded areas. Horizontal dashed lines represent the performance using the full training set. Models using AMPERE++ as the source domain consistently yield better F1 scores than others and non-TL models.

model (**MLM**) prediction, which randomly selects 15% of the input tokens for prediction as done in Devlin et al. (2019); (2) **context-aware sentence perturbation** (**Context-Pert**), which packs each document into a sequence of sentences segmented by the `[CLS]` token, 20% of which are replaced by random sentences from other documents, another 20% shuffled within the same document, and the rest unchanged. The pre-training objective is to predict the perturbation type of each sentence. Results are in the middle part of Table 3, where MLM pre-training benefits all three domains. Context-Pert improves AMPERE++ even more, but negatively affects the other two domains.

Combining Inductive and Transductive TL. Moreover, we showcase that *adding self-supervised learning as an extra pre-training step for transductive TL further boosts performance.* From the lower half of Table 3, the pre-trained model uniformly improves over the standard transductive TL. Notably, *using target domain for pre-training leads to better results than using the source domain data*. This implies that better representation learning for target domain language is more effective than a stronger source domain model.

Effectiveness of TL in Low-Resource Setting. To quantitatively demonstrate how TL benefits low-resource target domains, we control the size of training data and conduct transductive TL for each domain. Fig. 4 plots the trends where training data varies from 0 to 5,000, incremented by 500. *Among all datasets, AMPERE++ yields the best transfer learning results as the source domain: Using less than half of the target training set, it allows to approach or exceed the fully trained models.* For other datasets, we observe mixed results when they are used as the source. In general, TL brings more improvements when less training data is used.

6.3 Active Learning Results

Comparisons of Acquisition Strategies. Fig. 5 plots the F1 scores for all strategies as discussed in §4 across 10 AL iterations. As expected, the performance gradually improves with more labeled data. The three model-based methods: **MAX-ENTROPY**, **BALD**, and **CORESET** generally attain better performance, suggesting the efficacy of common AL methods on argument relation understanding. The model-independent strategies yield competitive results. In particular, **DISC-MARKER** proves to be

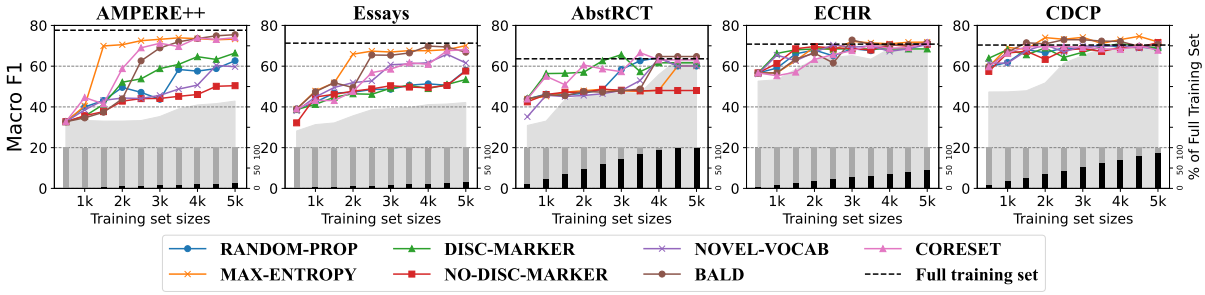

Figure 5: Active learning results using different acquisition methods in 10 iterations. Shaded areas stand for the **RANDOM-CTX** performance, which aligns with that in Figure 4. We show performance for three model-independent strategies, **DISC-MARKER**, **NOVEL-VOCAB**, **NO-DISC-MARKER**, alongside three strong comparisons. The model-independent strategies yields significantly better results than random sampling. On AbstRCT, ECHR, and CDCP, **DISC-MARKER** achieves better or competitive performance than **MAX-ENTROPY** and **BALD**. To better visualize the performance difference, rescaled plots for ECHR and CDCP are in Appendix A.3.

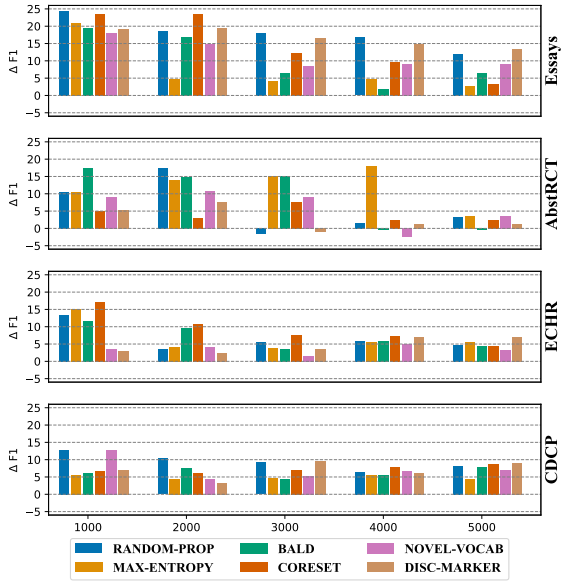

Figure 6: Improvements of macro F1 scores by adding TL to each AL strategy. AMPERE++ is used as the source domain for TL. We observe consistent gains across the board except for AbstRCT when the training samples are close to full. Generally, the improvements decline when more training samples are included.

a good selection heuristics for AMPERE++ and AbstRCT. Its relatively low scores on Essays is likely due to the abundance of discourse markers in this domain, so that random sampling would have similar effects. By contrast, avoiding discourse markers (**NO-DISC-MARKER**) tends to hurt performance. Notably, *without relying on any trained model, task-specific acquisition strategies can be effective for labeling argument relations*.

Warm-start Active Learning. Finally, we investigate the added benefits of transfer learning for major active learning systems. In each AL iteration, we warm-start the model with checkpoints trained from AMPERE++, and calculate the difference of F1 scores from the non-TL counterpart. Fig. 6 shows the results for five of the ten iterations. We observe improvements across the board, especially with small training data size. For AbstRCT, the TL warm-start either makes no difference or slightly hurts performance after 3,000 samples are available, whereas the **MAX-ENTROPY** method constantly benefits from warm-starting. Our findings suggest that TL is an effective add-on for early stage AL, benefiting different strategies uniformly.

7 Conclusion

We present a simple yet effective framework for argument structure extraction, based on a context-aware Transformer model that outperforms strong comparisons on five distinct domains, including our newly annotated dataset on peer reviews. We further investigate two complementary frameworks based on transfer learning and active learning to tackle the data scarcity issue. Based on our extensive experiments, transfer learning from our newly annotated AMPERE++ dataset and self-supervised pre-training consistently yield better performance. Our model-independent strategies approach popular model-based active learning methods.

Acknowledgements

This research is supported in part by National Science Foundation through Grant IIS-2100885. We thank four anonymous reviewers for their valuable suggestions on various aspects of this work.

References

Pablo Accuosto and Horacio Saggion. 2019. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In *Proceedings of the 6th Workshop on Argument Mining*, pages 41–51, Florence, Italy. Association for Computational Linguistics.

Charu C Aggarwal, Xiangnan Kong, Quanquan Gu, Jiawei Han, and S Yu Philip. 2014. Active learning: A survey. In *Data Classification: Algorithms and Applications*, pages 571–605. CRC Press.

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Zalán Bodó, Zsolt Minier, and Lehel Csató. 2011. Active learning with clustering. In *Active Learning and Experimental Design workshop In conjunction with AISTATS 2010*, volume 16 of *Proceedings of Machine Learning Research*, pages 127–139, Sardinia, Italy. PMLR.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-driven analysis. In *IJCAI*, volume 18, pages 5427–5433.

Claire Cardie, Cynthia Farina, Matt Rawding, and Adil Aijaz. 2008. An eRulemaking corpus: Identifying substantive issues in public comments. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. 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.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip Cohen, and Mark Johnson. 2018. Active learning for deep semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 43–48, Melbourne, Australia. Association for Computational Linguistics.

Aris Fergadis, Dimitris Pappas, Antonia Karamolegkou, and Haris Papageorgiou. 2021. Argumentation mining in scientific literature for sustainable development. In *Proceedings of the 8th Workshop on Argument Mining*, pages 100–111, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nancy Green. 2014. Towards creation of a corpus for argumentation mining the biomedical genetics research literature. In *Proceedings of the First Workshop on Argumentation Mining*, pages 11–18, Baltimore, Maryland. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Steven CH Hoi, Rong Jin, and Michael R Lyu. 2006. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. 2019. Argument mining for understanding peer reviews. 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 2131–2137, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 963–973, Vancouver, Canada. Association for Computational Linguistics.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. 2009. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2020. Adapting BERT to implicit discourse relation classification with a focus on discourse connectives. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France. European Language Resources Association.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

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*.

David Lowell, Zachary C. Lipton, and Byron C. Wallace. 2019. Practical obstacles to deploying active learning. 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 21–30, Hong Kong, China. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Tobias Mayer, Elena Cabrio, and Serena Villata. 2020. Transformer-based argument mining for healthcare applications. In *ECAI 2020 - 24th European Conference on Artificial Intelligence*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2108–2115. IOS Press.

Margot Mieskes and Andreas Stiegelmayr. 2018. Preparing data from psychotherapy for natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Huy Nguyen and Diane Litman. 2016. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany. Association for Computational Linguistics.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2019. Dissecting content and context in argumentative relation analysis. In *Proceedings of the 6th Workshop on Argument Mining*, pages 25–34, Florence, Italy. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Prakash Poudyal, Jaromir Savelka, Aagje Ieven, Marie Francine Moens, Teresa Goncalves, and Paulo Quaresma. 2020. ECHR: Legal corpus for argument mining. In *Proceedings of the 7th Workshop on Argument Mining*, pages 67–75, Online. Association for Computational Linguistics.

Patrick Saint-Dizier. 2018. A two-level approach to generate synthetic argumentation reports. *Argument & Computation*, 9(2):137–154.

Claudia Schulz, Christian M. Meyer, Jan Kiesewetter, Michael Sailer, Elisabeth Bauer, Martin R. Fischer, Frank Fischer, and Iryna Gurevych. 2019. Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772, Florence, Italy. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*.

Burr Settles. 2009. Active learning literature survey.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

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

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Katrin Tomanek and Udo Hahn. 2009. Reducing class imbalance during active learning for named entity annotation. In *Proceedings of the fifth international conference on Knowledge capture*, pages 105–112.

Simon Tong and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66.

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. Curran Associates, Inc.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*.

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.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start active learning through self-supervised language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

A Model and General Details

A.1 Discourse Markers

In §4.2 of the main paper we introduce a **DISC-MARKER** based acquisition method for active learning. The matching statistics of the 18 discourse markers are shown in Fig. 8. We break down the count based on whether a proposition is the `head` or `tail` of any relation. As expected, certain discourse markers such as "because", "but", and "due to" likely indicate a `tail` proposition, whereas "therefore", "thus" tend to be found in `head` propositions.

A.2 SVM Comparison

In Table 4, we describe the full feature set used in the SVM comparison model in § 6.1 of the main paper. These features are adapted from Table 10 of Stab and Gurevych (2014). The indicators are from their Table B.1 in the Appendix.

For hyper-parameter search, we tune the regularization coefficient C over values $\{0.1, 0.5, 1.0, 10.0\}$. The best performing model (macro-F1) on validation set is used for evaluation.

Group	Description
Lexical	Binary lemmatized unigram of the head and tail propositions (top 500 frequent ones are considered)
Syntactic	Binary POS features of head and tail propoisitions
Structural	Number of tokens of head and tail; Number of propositions between source and tail; head presents before tail; tail presents before head
Indicator	Indicator type present in head or tail; indicator type present between head and tail
ShNo	Shared nouns between head and tail propositions (number and binary)

Table 4: Features used for SVM model.

A.3 Active Learning Results

In Fig. 5, we compare active learning methods over five datasets on the same 0-80 scale. Results of different strategies fall in tight ranges for ECHR and CDCP. For better visualization, we show the same figure on a 50–80 scale in Fig. 7.

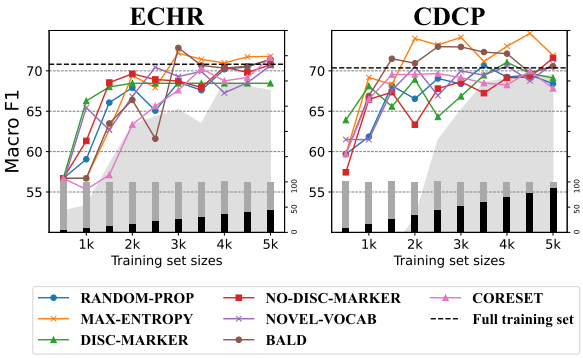

Figure 7: Active learning results for ECHR and CDCP on 50–80 scale. The scores are the same as the rightmost two plots in Figure 5.

B AMPERE++ Annotation

To annotate argument relations over the AMPERE (Hua et al., 2019) dataset, we hire three proficient English speakers who are US-based college students. The first author serve as the judge to resolve disagreements.The detailed guidelines are shown in Table 6. Throughout the annotation, we identify difficult cases and summarize representative ones in Table 5.

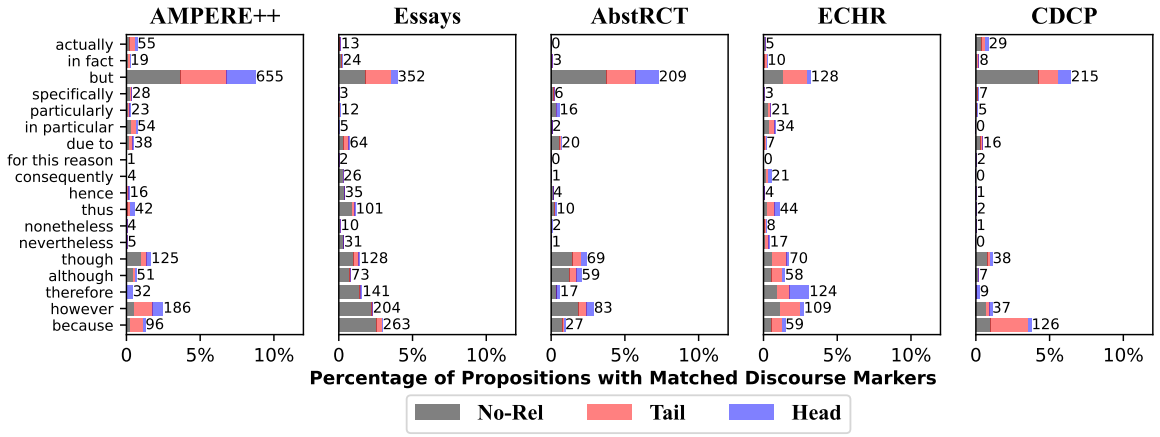

Figure 8: The distribution of matched discourse markers in each dataset. We indicate the raw count next to each bar. Propositions that are `tail` or `head` of any relation are highlighted in colors. Overall, about 10–20% of the propositions contain at least one discourse marker. Certain discourse markers correlate well with the existence of argument relations. For instance, "because", "due to", "however" are more likely to be found in `tail`; "therefore", "thus" tend to appear in `head`.

Tail	Only macro-average F-scores are reported.
Head	Please present micro-average scores as well
Label	`support`
Tail	Fig 3: This could really be drawn considerably better
Head	Make the dots bigger and their colors more distinct.
Label	`support`
Tail	Fig 4. right looks like a reward signal.
Head	but is labelled Proportion correct.
Label	`attack`
Tail	This idea is not novel
Head	In the first part of the paper (Section 2) the authors propose to use the optimal transport distance . . . as the objective for GAN optimization.
Label	`attack`
Tail	Then, the difference is crystal clear.
Head	The difference between Figure 1, 4, and 6 could be clarified.
Label	`no-rel`
Tail	The discussion following Corollary 1 suggests that $\sum_i \hat{v}_{T,i}^{1/2}$ might be much smaller than d G_∞.
Head	but we should always expect it to be at least a constant,
Label	`no-rel`

Table 5: Representative challenging examples during argument relation annotation on AMPERE++.

General Instruction
In the following studies, you will read a total of 400 peer reviews collected from the ICLR-2018 conference. The annotation is carried out in 20 rounds. In each round, you will independently annotate 20 reviews and upload to the server. All annotators will meet and discuss the disagreements. Another judge will resolve the cases and add it to the pool of samples for future reference.

Annotation Schema
Each review document is already segmented into chunks of argumentative discourse units (ADU), which is the basis for relation annotation. Prior work has provided labels for types of these ADUs:

EVALUATION: Subjective statements, often containing qualitative judgement.
REQUEST: Statements requesting a course of action.
FACT: Objective information of the paper or commonsense knowledge.
REFERENCE: Citations or URLs.
QUOTE: Quoations from the paper.
NON-ARG: Non-argumentative statements.

Please first read the entire review. Then, from the beginning of the document, start annotating support and attack relations. We consider a support relation holds from proposition `A` to proposition `B` if and only if the validity of `B` can be undermined without `A`, or `A` presents concrete examples to generalize `B`. For example, "*It is unclear which hacks are the method generally.*" is supported by "*Because the method is only evaluated in one environment.*".

We consider an attack relation holds from proposition `A` to proposition `B` if and only if `A` contrasts or questions `B`'s stance. For example, "*The authors mentioned that the grammar in general is not context free.*" is attacked by "*But the grammar is clearly context-free.*"

Both the support and attack relations can be implicit or explicit. Explicit relations are indicated by discourse markers, whereas implicit relations require inference from the context. For example, "*In particular, how does the variational posterior change as a result of the hierarchical prior?*" implicitly supports "*It's not clear as to why this approach is beneficial*". Because the question instantiates the "unclear" claim regarding the approach.

Special Cases
Please enforce the following constraints:

1. The factual propositions (i.e., FACT, REFERENCE, QUOTE) cannot be supported by any subjective propositions (i.e., EVALUATION, REQUEST).
2. One proposition can support or attack at most one proposition.
3. Chain support does not need to be explicitly annotated. For instance, if `A` supports `B`, `B` supports `C`, then `A` supports `C` does not need annotation.

Table 6: Argumentative relation annotation guideline for AMPERE++.