Discourse-based Argument Segmentation and Annotation

Ekaterina Saveleva and **Volha Petukhova** and **Marius Mosbach** and **Dietrich Klakow**
Spoken Language Systems Group, Saarland Informatics Campus
Saarland University, Saarbrücken, Germany
{esaveleva,vpetukhova,mmosbach,dklakow}@lsv.uni-saarland.de

Abstract

The paper presents a discourse-based approach to the analysis of argumentative texts based on the assumption that the coherence of a text should capture argumentation structure. Therefore, existing discourse analysis tools can be successfully applied for argument segmentation and annotation tasks. We tested widely used Penn Discourse Tree Bank parser (Lin et al., 2010) and the state-of-the-art neural network NeuralEDUSeg (Wang et al., 2018) and XLNet (Yang et al., 2019) models on discourse segmentation and discourse relation recognition tasks. The two-stage approach outperformed the PDTB parser by broad margin, i.e. the best achieved F1 scores of 21.2% for PDTB parser vs 66.37% for NeuralEDUSeg and XLNet models. Neural network models were fine-tuned and evaluated on the argumentative corpus showing a promising accuracy of 60.22%. The complete argument structures were reconstructed for further argumentation mining tasks. The reference Dagstuhl argumentative corpus containing 2,222 elementary discourse unit pairs annotated with the top-level and fine-grained PDTB relations will be released to the research community.

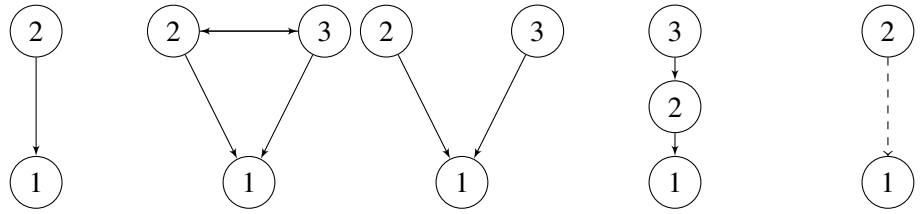
1 Introduction

Enormous and ever growing digital content provides information where opinions, sentiment and arguments can be identified and analysed. For example, news and social media content is searched to filter or weight the validity of statements (Rowe and Butters, 2009), to identify the presence of fake news and false claims (Popat et al., 2018), to analyse opinions in public discussions (Murakami and Raymond, 2010), to detect opinion manipulation (Cambria et al., 2010), to predict consumers sentiment (Bai, 2011), to study citizen engagement (Purpura et al., 2008), and to recognize stance in political online debates (Somasundaran and Wiebe, 2010; Walker et al., 2012). Arguments from legal (Moens et al., 2007), financial (Hogenboom et al., 2010) or medical (Sanchez Graillet and Cimiano, 2019) documents are extracted to support professional decision-making. Natural argumentation is the focus of numerous educational scenarios assessing student's essays quality (Stab and Gurevych, 2017) and training argumentation and debate skills (Ashley et al., 2007; Petukhova et al., 2017). Automatic extraction and analysis of arguments from heterogeneous data is one of the important tasks of *argumentation mining* which aims to provide structured data for computational models of argument and reasoning engines (Lippi and Torroni, 2016).

While for some applications, an argument can be considered as an atomic entity without internal structure, for others defining its structure becomes crucial. For example, to recognize the speaker 'stance'[1] in online debates, the whole post can be acknowledged as an argument in 'favour' or 'against' a certain motion. An argument is, therefore, analysed given the other supporting or attacking arguments (Dung, 1995). Other argumentation mining tasks require structured argumentation models, e.g. tasks that aim at understanding and emulation of human inference, investigating patterns of reasoning, and tasks that focus on extraction and validity assessment of arguments.

Identification and classification of argument components are rather challenging tasks (Aharoni et al., 2014). The argument definition, the description of elementary units and building blocks of an argument, relations between and inside these units, the argument structures and argumentation schemes are still under debate. A simple argument structure is often considered as consisting of *a claim* that is supported by *evidence* (Mochales and Moens, 2011; Aharoni et al., 2014). A claim is an assertion that the argument aims to prove, i.e. a claim is a *conclusion* whose merit must be estab-

[1] Stance is defined as an overall position held by a person towards an idea or attitude (Somasundaran and Wiebe, 2009).



(i) basic argument (ii) linked support (iii) multiple support (iv) serial support (v) example support

Figure 1: Basic support relations and complex formations suggested by Peldszus and Stede (2013).

lished. Evidence presents (a set of) proposition(-s) which provide grounds for drawing the conclusion.

Automatic recognition of relevant semantic units involves two tasks: (1) *segmentation* of a text into meaningful units; and (2) *annotation* of these units capturing (part of) their meaning. Many argumentation mining studies assume that the boundaries of the argument components have been previously detected by other means, thus they focus on the classification task (Stab and Gurevych, 2014; Eckle-Kohler et al., 2015). Other consider segmentation as a sub-task and perform both segmentation and classification (Levy et al., 2014; Rinott et al., 2015).

We define *argument structure recognition* to involve: (1) segmentation of a text into elementary argumentative units assuming that they correspond to elementary discourse units; (2) discourse relation detection between them; (3) classification of the identified relations; (4) classification of the identified argumentative units based on the classified discourse relations; and (5) argument completion – reconstruction of implicit units to achieve a complete argument structure, see also (Peldszus and Stede, 2013). In this study we evaluate state-of-the-art discourse parsers and machine learning models on automatic segmentation and discourse relation classification tasks, and then apply them to extract arguments from argumentative texts.

The rest of the paper is structured as follows. In Section 2, we discuss related work concerning argument structure recognition. Section 3 presents established discourse theories as theoretical and empirical framework for argument analysis and argumentation modelling. The connection to the existing ISO 24617-8 standard for discourse relation annotation is made. Section 4 discusses the performed experiments elaborating on the datasets, tools and outcomes. Section 5 summarizes the results and outlines the future research.

2 Related Work

Peldszus and Stede (2013) defined Argumentative Discourse Units (ADUs) as text segments corresponding to propositions that are argumentatively relevant and have their own argumentative function. ADUs reflect different ways to support a claim (Fig. 1), e.g. with the *basic* argument configuration consisting of a conclusion supported by exactly one premise, as in example (1) below. If there are multiple premises supporting a conclusion together, the structure is called *linked support* as in (2). Multiple premises which support the conclusion independently form a *multiple support* as in (3). *Serial* support links arguments to the conclusion where an argument contributes to further development of an already given argument (4). Peldszus and Stede (2013) consider the example shown in (5) to be a special form of support.

(1) [Books are better than TV.]$_1$ [Books enlighten the soul.]$_2$

(2) [Books are better than TV.]$_1$ [Books enlighten the soul.]$_2$ [They change your perspective on life]$_3$

(3) [Books are better than TV.]$_1$ [1. Books don't ruin your eyes like TV does.]$_2$ [2.Books allow your brain to imagine]$_3$ [3.Reading books can help you with spelling.]$_4$ [4.Reading books can help you write better.]$_5$

(4) [Gay marriage is wrong.]$_1$ [In fact, we would all become extinct,$_2$] [because without one man and one woman]$_3$ [there would be no reproduction.]$_4$

(5) [Personal pursuit is better than advancing the common good.]$_1$ [I need to think about me first, success and then think of others.]$_2$

Since not every text is argumentative and, therefore, subjected to an argumentative analysis, identification of its type can be considered as a preliminary step, and together with the topic context may provide valuable information for the argument component identification. Levy et al. (2014) introduced the notion of a *context-dependent claim* – a general concise statement that directly supports or contests a given topic. Rinott et al. (2015) detect *context-dependent evidence* – text segments that directly support a claim in the context of a given topic. Contextual information has served as an important source for argument component identification in

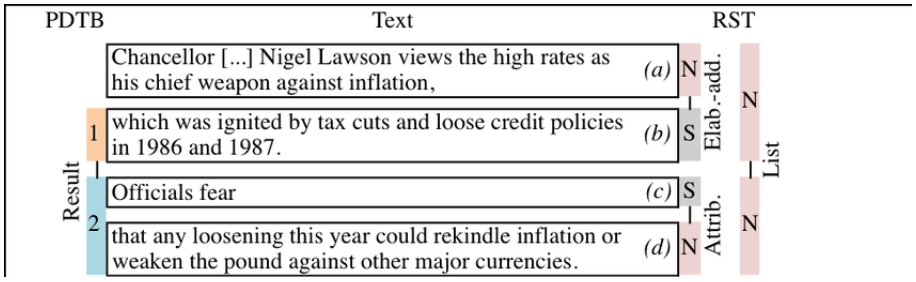


Figure 2: PDTB and RST-DT annotations for a *WSJ 1172* paragraph (Demberg et al., 2019), where 1 refers to Arg1 and 2 to Arg2 in PDTB; N stands for Nucleus and S for Satellite in RST; and (a-d) are RST-DT's Elementary Discourse Units.

Kuribayashi et al. (2018); Opitz and Frank (2019); Aker et al. (2017); Shnarch et al. (2018).

Mining arguments from diverse corpora based on topic can pose certain problems. A well-established topic is not always easy to determine or a text can cover several topics and the discussion can shift between them throughout the entire text. Lippi and Torroni (2015) proposed a method for *context-independent claim detection*. The approach relies on the assumption that argumentative sentences share the structure independently of the addressed topic. This technique was successfully applied for legal texts (Lippi et al., 2015), clinical trials (Mayer et al., 2018) and social media (Liga, 2019).

Cross-domain approach to the argumentation mining has been explored in a number of studies. Rosenthal and McKeown (2012) detect claims from two different data sets, LiveJournal and Wikipedia. Al Khatib et al. (2016) experimented with a wider range of text types and topics addressing politics, culture, religion, sport, economy, and health. Deep learning techniques were applied in cross-domain and multi-task learning scenarios (Eger et al., 2017; Daxenberger et al., 2017; Stab et al., 2018; Schulz et al., 2018; Morio and Fujita, 2019; Mensonides et al., 2019; Wambsganss et al., 2020).

Argument structure is often viewed through the prism of discourse theory and ADU components are defined based on discourse units which proves that argumentation and discourse characteristics, and these structures are closely related. Peldszus and Stede (2016) explored the mapping between discourse and argument(-ation) structures based on the Rhetorical Structure Theory (RST, Mann and Thompson (1988)) and those of Segmented Discourse Representation Theory (SDRT, Lascarides and Asher (2008)). Stede et al. (2016) assesses the role of discourse parsing features for argumentation structure prediction. Cabrio et al. (2013) and Hewett et al. (2019) translated the general simple argument structure into several discourse-based schemes to perform analysis and evaluation of natural language arguments, see also (Teufel et al., 1999; Palau and Moens, 2009; Petukhova et al., 2017). Eckle-Kohler et al. (2015) assessed the role of discourse markers for claims and evidence detection. In Hofmockel et al. (2017), the impact of the genre on different realizations of discourse relations is evaluated. Green (2018) applied the genre-based approach to scientific (e.g. biological/biomedical) texts.

3 Discourse Analysis

Discourse theory aims at explaining the coherence of a text. Its central notion is *coherence*, also called *rhetorical* or *discourse relation* - a semantic or pragmatic relation between two adjacent text spans. Even though text coherence and argumentation structure are not identical, discourse structure can reveal new unexplored properties of argumentation. Bridging from discourse to argumentation, Peldszus and Stede (2013) chooses the RST framework where all parts of a text are involved into a discourse structure and organized as a tree, with Elementary Discourse Units (EDUs) as leaves. An EDU is a minimal building block of a discourse tree which typically corresponds to a clause (Carlson and Marcu, 2001).[2] RST specifies how EDUs and larger units are connected, where some text spans are more important than the others, i.e. *nucleus* or multiple *nuclea* are the central part of a relation in the text supported by a *satellite*. The corresponding RST tagset contains 78 discourse relations which can be grouped into 16 classes sharing one type of rhetorical meaning.

Another influential discourse analysis frame-

[2]Other competing hypotheses take an EDU to be a prosodic unit, a dialogue turn, a sentence, an intentionally defined discourse segment (e.g. utterance) or the contextually indexed representation of information.

work is defined within Penn Discourse Tree Bank (PDTB, Prasad et al. (2005)). PDTB does not make strong assumptions about the overall structure of a text and does not suggest what kinds of high-level structures may be created from the annotated low-level relations and arguments. The PDTB analysis is focused on the discourse relation between two text segments called `Arg1` and `Arg2` which can be treated as EDUs. PDTB accounts for the lexical items that can signal discourse relations – discourse connectives. In the case of *explicit* connectives, `Arg2` is the argument to which the connective is syntactically bound, and `Arg1` is the other argument. In the case of relations between adjacent sentences, `Arg1` and `Arg2` reflect the linear order of the arguments, with `Arg1` before `Arg2`. PDTB does not constrain an EDU to be a single clause or single sentence, however, the framework follows a minimality principle requiring an argument to contain the minimal amount of information needed to interpret the relation successfully. The PDTB annotation scheme forms the basis of the ISO DR-Core (ISO 24617-8) discourse relations annotation standard (Bunt and Prasad, 2016).

Even though RST and PDTB annotation frameworks make different assumptions about the discourse structure and define different sets of relations, Demberg et al. (2019) suggest an automatic alignment of their relations and evaluates the mapping discrepancies. Figure 2 compares PDTB and the RST Treebank (RST-DT, Carlson et al. (2003)) annotations of the WSJ-1172 paragraph of Penn Tree Bank (PTB, Marcus et al. (1993)).

Discourse analysis within both annotation frameworks includes (1) segmentation of the text into EDUs; and (2) the recognition of discourse relations between these units. Discourse parsers typically perform both tasks. For example, Lin et al. (2010) designed a full parser to perform the PDTB annotations. The system first identifies discourse connectives, label the corresponding `Arg1` and `Arg2` spans and assign an `Explicit` relation. If no connective was identified, the system classifies the statement pair as having one of the other relation types, i.e. `Implicit, EntRel, AltLex, NoRel`.

Wang and Lan (2015) extended the parser with extractors for `Arg1`, `Arg2` and `Non-EntRel` relations. Qin et al. (2016) improved recognition of the implicit relations. Recent works explore deep learning techniques which use architectures for multi-task learning (Liu et al., 2016; Lan et al., 2017; Van Ngo et al., 2019) or adversarial neural networks (Qin et al., 2017; Huang and Li, 2019).

While many studies focus exclusively on the discourse relation recognition assuming that the text is already pre-segmented, others also consider discourse segmentation task. Early generation segmenters were rule-based systems (LeThanh et al., 2004; Tofiloski et al., 2009), whereas more recent approaches view this task as sequence labeling problem and use deep learning (Hernault et al., 2010; Bach et al., 2012; Wang et al., 2018). Multilingual discourse segmentation is addressed in (Braud et al., 2017; Muller et al., 2019; Desai et al., 2020).

The presented study concerns both segmentation and classification tasks assessing the performance of the state-of-the-art tools on the argumentative corpus. Design and results are reported in the next Section.

4 Experimental Design

We conducted the following experiments: (1) evaluating the quality of the existing full discourse parsers on EDUs segmentation and relation classification tasks; (2) two-stage discourse segmentation and relation annotation; (3) application and evaluation of the best performing model to identify and classify argument components in the argumentative corpus; and (4) completion of argument structure by reconstructing implicit claims. Figure 3 shows the experimental workflow.

4.1 Datasets

There are two corpora used in this study: *Penn Discourse Treebank 2.0* (PDTB 2.0, Prasad et al. (2008))[3] – a large scale corpus annotated with information related to discourse structure and discourse semantics, and *Dagstuhl15512 ArgQuality* (Wachsmuth et al., 2017) – a corpus of segmented arguments annotated with argument quality scores.

PDTB 2.0 consists of 2,159 articles from Wall Street Journal (WSJ) divided into 25 sections. In total, there are 40600 discourse unit pairs annotated with different relations. We provide a list of relations and their distribution in the Appendix.

[3]PDTB 2.0 is an extended version of the PDTB 1.0 corpus, where extensions concern annotations of implicit relations for the entire corpus, senses of all connectives and attribution of object type, scopal polarity and determinacy. Thus, for the purpose of this study, differences between PDTB 1.0 and PDTB 2.0 are not relevant.

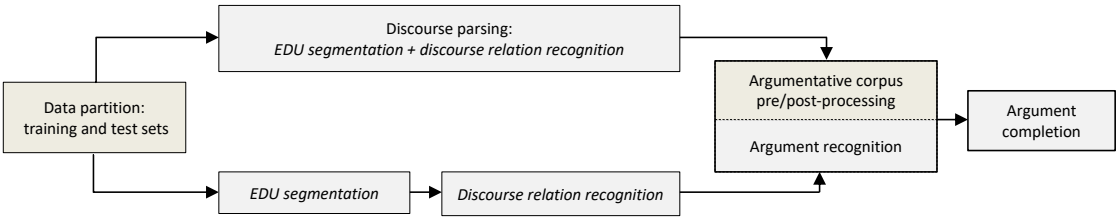


Figure 3: Experimental workflow for the argument structure recognition.

Dagstuhl15512 ArgQuality (Wachsmuth et al., 2017) is a collection of argumentative texts from the *UKPConvArgRank* dataset (Habernal and Gurevych, 2016) consisting of debate portal arguments *for* and *against* stances on 16 topics. The *UKPConvArgRank* dataset was developed to predict the convincingness of arguments, so that each argument pair is rated as more or less convincing. For *Dagstuhl15512 ArgQuality*, texts on each topic containing the five top- and five bottom-ranked arguments were selected and annotated across three core quality dimensions: argument cogency, argument effectiveness and argument reasonableness, and several sub-criteria (15 in total).

4.2 Discourse Analysis Tools Assessment

4.2.1 Discourse Parsing

One of the widely used tools for discourse processing is the PDTB parser developed by Lin et al. (2010). It is trained on sections 02-21 of Penn Discourse Tree Bank (PDTB 1.0, Prasad et al. (2005)) for text span identification and relation classification. For our purposes, spans for both `Arg1` and `Arg2` need to correspond exactly or partially to the PDTB 2.0 reference segments. Moreover, the relation between EDUs should be correctly classified. We evaluated the parser performance on the full PDTB 2.0 corpus. Table 1 summarizes parser performance in terms of F1 scores. The gold standard parsing and EDUs boundaries with error propagation setting ($GS + EP$) refers to a clean, per-component evaluation. In the automatic parsing and EDUs boundaries with error propagation scenario ($Auto + EP$), end-to-end automated parsing of the unseen data is performed. In the later setting, F1 scores of 38.18% and 20.64% were achieved for partial and exact match, respectively. A large portion of the misclassified cases belong to the `Non-Explicit` classes, as implicit discourse relations are more difficult to classify. The bottom part of Table 1 reports F1 scores obtained on the EDU span identification and on the joint segmentation and classification tasks on the entire PDTB 2.0 corpus.

Experimental setting	**F1 score** (%)
GS + EP (partial match)	46.80*
Auto + EP (partial match)	38.18*
GS + EP (exact match)	33.00*
Auto + EP (exact match)	20.64*
EDU span identification	22.61**
EDU span identification & relation recognition	21.20**

Table 1: Performance (F1 scores) of the PDTB parser developed by Lin et al. (2010) on various tasks. * evaluation performed on the section 23 of the PDTB 2.0 corpus; ** evaluation performed on the on full PDTB 2.0 corpus.

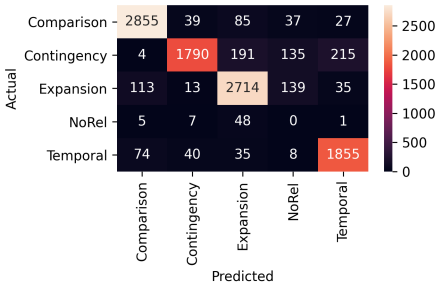


Figure 4: Confusion matrix for L1 relation classification with the PDTB parser.

Similarly to Hewett et al. (2019), we observed that the parser failed to identify many spans correctly. In case of the correct span identification, relation classification was reasonably accurate. Figure 4 shows the confusion matrix for the top-level (L1) relations between the correctly identified pairs of `Arg1` and `Arg2`. We concluded that the parser generally tends to assign a relation between the majority of EDU spans misclassifying `NoRel` instances.

4.2.2 Discourse Segmentation and Relation Recognition

As shown in the parser evaluation experiments, EDU segmentation is a crucial step in discourse analysis. Since the PDTB parser failed to show satisfactory segmentation performance, we tested state-of-the-art neural network model on the reference PDTB annotation, i.e. the BiLSTM-CRF

EDU segmentation			*PDTB Relation recognition*			
Statistics		**F1 score** (%)	**Statistics**			**Accuracy** (%)
			# Classes	Training set	Test set	
total segments	123780		2 classes	62172	4655	88.86
exact matches	13420 (10.84%)	68.55	5 classes	19145	4655	66.37
partial matches	56847 (45.92%)		10 classes	12070	4471	53.64

Table 2: Segmentation performance (F1 scores) with an overview of the exact and partial matches processed with NeuralEDUSeg (Wang et al., 2018) of the PDTB 2.0 corpus; and relation recognition accuracy for different classification scenarios applying the XLNet model (Yang et al., 2019) on PDTB 2.0 data.

based model NeuralEDUSeg developed by Wang et al. (2018). In this experiment, a unit is acknowledged to be correctly segmented if it partially or fully corresponds to one of the reference PDTB segments. Table 2 reports the number of exact and partial matches. Segmentation performance achieves 68.55% of F1 score. Our results show that NeuralEDUSeg significantly outperforms the PDTB parser (compare with Table 1). While the number of exact matches is still rather low (10.84%), we observed a relatively high number of identified partial matches (45.92%). The fact that most matches coincide with the reference segmentation only partially can be explained by the fact that NeuralEDUSeg is originally trained on the RST-DT corpus which follows different segmentation principles (consider Figure 2 again). Minimal RST-DT units tend to be shorter than those of the PDTB. For example, compare the NeuralEDUSeg [segment]$_1$ with the PDTB [segment]$_2$ illustrated in (6):

(6) a) [Woolworth said]$_1$ [Woolworth said it expects to expand usage of the MCI services as it adds about 6000 business locations over the next few years]$_2$
b) [The derivative markets remained active]$_1$ [The derivative markets remained active as one new issue was priced]$_2$

Deep learning models show promising results on discourse relation recognition task. Kim et al. (2020) demonstrated that the XLNet-large model of Yang et al. (2019) achieved the best results on implicit discourse relation recognition significantly outperforming BERT- (Nie et al., 2019) and ELMO-based (Bai et al., 2019) discourse relations models.

We performed a series of experiments on fine-tuning XLNet for the discourse relation recognition task. We first conducted a binary classification to establish whether is any relation between the identified units, i.e. the model discriminates between `Rel` class (includes any type of discourse relations) and `NoRel` comprising the `EntRel` and `NoRel` types. Secondly, we performed five-class top-level (L1) and ten-class fine-grained (L2) relations classification. The following five classes were used for the second experiment: *Expansion, Conjunction, Comparison, Contingency, Temporal, NoRel*. The ten-class experiment exploited the classes listed below: *Expansion.Conjunction, Expansion.Restatement, Expansion.Instantiation, Temporal.Synchrony, Temporal.Asynchronous, Contingency.Cause, Contingency.Condition, Comparison.Contrast, Comparison.Concession*. See Appendix for the class distribution. Classes with less than 500 training instances were excluded. The training set comprised sections 0-21 of the PDTB 2.0 corpus; sections 22-24 served as the test set. Since classes were not balanced in all classification settings, we performed *re-sampling* procedure: *up-sampling* of the under-represented `NoRel` class in binary classification by adding synthetic samples combining random EDUs from different textual units; and *down-sampling* the majority classes in the multi-class settings. The right part of Table 2 presents the final training and test data partitions for each classification scenario.

For the training and evaluation procedure, we fine-tuned each encoder model following the suggestions of Mosbach et al. (2021) and trained for 10 epochs using a learning rate of 0.00001 and a batch-size of eight. The results are summarized in Table 2, from which we can observe that accuracy drops with a higher number of classes to learn, from 88.85% for two classes to 53% for ten classes. We note that the results of our experiments differ from those reported by Kim et al. (2020) due to the differences in the approach and set of the classified relations. For instance, we were not focused on the distinction between implicit vs. explicit relation recognition. The goal was to assess how well the model predicts cases when a relation between two segments exists without focusing on how this relation is expressed. Moreover, we included `NoRel` instances into the classification, while they are typically discarded in other studies.

4.3 Discourse-based Analysis of an Argumentative Corpus

We applied the tested discourse analysis tools on *Dagstuhl15512 ArgQuality* corpus where we manually examined and corrected the model outputs. Respecting the PDTB minimality principle, we combined or split relevant text units depending on the amount of information required to interpret the relation between the segments correctly. We conduct a detailed error analysis and discuss some representative cases below.

We encountered many examples where a single unit does not contain substantial semantic information and has to be combined with the adjacent segment(-s) as illustrated in (7):[4]

(7) [A law] [requiring separate schools and public accommodations for homosexual people would violate] ["separate but equal"] → [A law requiring separate schools and public accommodations for homosexual people would violate "separate but equal"]

We considered modal constructions such as *I think, I believe, I am sure, Maybe, I highly doubt* as in (8), infinitive constructions (9), participle constructions (10) and relative clauses (11) as not forming an EDU on their own and therefore not having any discourse relation to the neighbouring EDU(-s). Relevant segments are merged.

(8) [I believe] [it should not be done] [just to discipline a child.] → [I believe [it should not be done just to discipline a child.]

(9) [Congress have no power] [to pass a legislation] [forcing religious institutions about marriage.] → [Congress have no power to pass a legislation forcing religious institutions about marriage.]

(10) [it doesn't break the Separation between Church and State] [ruled by the Supreme Court.] → [it does n't break the Separation between Church and State ruled by the Supreme Court.]

(11) [It would be hard for me to turn in the one] [I love.] → [It would be hard for me to turn in the one I love.]
[Yes, if the person] [I loved] → [Yes, if the person I loved]

We also encountered a few cases where the segment identified by the parser can be split into several EDUs as in (12):

(12) [So, many countries depends on scientists.most of employees in every country] [is Indians.], [and still be successful. Take myself for example;] → [So, many countries depends on scientists.] [most of employees in every country is Indians.]

EDU segmentation		*PDTB relation recognition*	
Match type	**F1 score** (%)	**# Classes**	**Accuracy** (%)
exact match	47.94	5 classes	60.22
partial match	79.83	10 classes	50.48

Table 3: Performance on EDU segmentation task applying NeuralEDUSeg model Wang et al. (2018) on the Dagstuhl corpus in terms of F1 scores (in %); and accuracy scores (in %) for 5- and 10 class discourse relation classification on the DagStuhl corpus with the fine-tuned XLNet-large model.

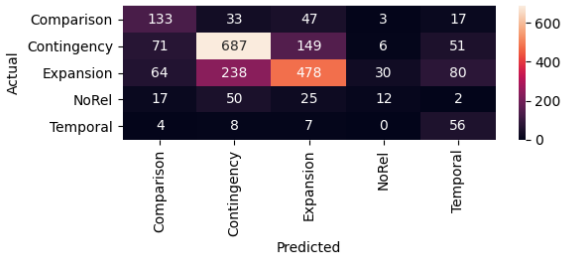


Figure 5: Confusion matrix for 5-class discourse relation classification task on Dagstuhl15512 ArgQuality.

Table 3 reports the performance of the NeuralEDUSeg model evaluated on the manually segmented argumentative Dagstuhl corpus using the reference segmentation.

As the next step, the identified EDUs were used to classify discourse relation between them applying XLNet (Yang et al., 2019). For this, pairs of adjacent segments were constructed and annotated with the PDTB discourse relations. We focused on ten classes mentioned above (the distribution is provided in the Appendix).

Wachsmuth et al. (2017) notes that some argument components, most often a claim, can be implicit. Consider an example in (13) below. An argument is not complete without the claim and cannot be used for further argumentation mining tasks. Therefore, we reconstructed a claim for every topic in the corpus, i.e. either 'for' or 'against' stance it may present. The reconstructed claim is a simple sentence which correspond to a single EDU. Subsequently, the reconstructed claims were used to created EDUs pairs for discourse relation classification.

(13) (a) The question is: who has the right to prohibit it? Government? Why would there be any pressing need at all for the state to outlaw pornography? Look at Europe- they're cool with pretty much everything. I don't see

[4]Here and in the following examples, a text span in the square brackets corresponds to an EDU obtained with a neural discourse segmenter; the manually corrected version is given after the arrow sign →.

any moral depravity in Europe, do you? (implicit claim: Pornography is not wrong.)
(b) Books will be always great whatever the new technological developments emerges, books has its fixed place in every humans heart. (implicit claim: Books are better than TV.)

EDUs pairs were built considering non-adjacent text units connected by a discourse relation. Most frequently, a claim may be connected to segments representing various types of evidence at different support levels as in (14):

(14) [Advancing the common good is better than personal pursuit.] [I think common good is better than personal pursuit]
[Advancing the common good is better than personal pursuit.] [When people help each other out its more likely that everything comes out great.]
[Advancing the common good is better than personal pursuit.] [Yes personal pursuit is important]

The resulting *Dagstuhl* corpus annotated with discourse relations contains the same number of 304 arguments as the original one which are segmented into 2,222 EDUs pairs. The XLNet-large model, initially trained and fine-tuned on the PDTB 2.0 dataset, was evaluated on Dagstuhl, see Table 3 for the performance overview. Figure 5 presents the confusion matrix for the 5-class relation classification task. We observed that many relations are correctly classified even in the absence of discourse connectives on which the model relies. Consider the following classification output:

(15) Creationism tries to sneak the supernatural as a scientific explanation. `Expansion.Restatement` This is called pseudo - science.

So a lousy father is better than none. `Comparison.Concession` (that is of course assuming that he is not abusive in any way)

Books enlighten the soul. `Expansion.Conjunction` Books don't destroy the morals of children.

and the big corporations like Dasani and Nestle would loose millions of dollars. `Contingency.Cause` It would hurt the economy severely .

Physical education does absolutely nothing for the children 's health and/or lifestyle . `Expansion.Instantiation` Let me describe my PE experience. Throughout my public education career , PE has been mandatory for each year.

I think common good is better than personal pursuit `Comparison.Contrast` Yes personal pursuit is important.

it wouldn't be so easily for you to become fat `Contingency.Condition` (of course you would also need to keep a balanced diet)

To summarize, the evaluated discourse processing tools showed a reasonable segmentation (F1 score ranging from 47.94% for exact match to 79.83% for partial match) and discourse relation recognition (accuracy ranging from 50.48% to 60.22%) performance on argumentative data. Thus, they can be applied in argument structure recognition and reconstruction tasks.

5 Conclusions and Future Work

The presented study reviewed discourse-based approaches to argumentative discourse analysis. We evaluated three widely used tools on argument segmentation and annotation tasks, namely, a rule-based PDTB full parser (Lin et al., 2010), a BiLSTM-CRF model for discourse units segmentation (Wang et al., 2018) and an XLNet based discourse relations classifier (Yang et al., 2019). Our experiments demonstrated that the PDTB parser achieved an F1 score of 22.61% on the span identification and 21.20% on the joint span identification and relation recognition tasks. This performance has been considered unsatisfactory for further use. Deep learning models, in contrast, showed significantly better performance: F1 scores ranging from 47.94% to 79.83% were achieved on the segmentation task, and accuracy of 60.22% and 50.48% for top-level and fine-grained discourse relation classification, respectively.

We successfully applied the best performing models to segment and annotate the argumentative corpus *Dagstuhl15512 ArgQuality* and conducted the detailed error analysis. The obtained argumentative discourse units were manually corrected and annotated with the fine-grained PDTB discourse relations. This corpus contains 2,222 annotated unit pairs and presents a valuable resource for further argumentation mining studies and will be released to the community.

The obtained results opened up many interesting prospects for future research. For example, various argumentation schemes can be reconstructed based on the proposed approach, and evaluated within numerous contexts and domains. Argument and argumentation quality can be assessed and robust reasoning engines designed.

Acknowledgments

The research reported in this paper was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 232722074 SFB 1102.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first workshop on argumentation mining*, pages 64–68.

Ahmet Aker, Alfred Sliwa, Yuan Ma, Ruishen Lui, Niravkumar Borad, Seyedeh Ziyaei, and Mina Ghobadi. 2017. What works and what does not: Classifier and feature analysis for argument mining. In *Proceedings of the 4th Workshop on Argument Mining*, pages 91–96.

Khalid Al Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1395–1404.

Kevin Ashley, Niels Pinkwart, Collin Lynch, and Vincent Aleven. 2007. Learning by diagramming supreme court oral arguments. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 271–275.

Ngo Xuan Bach, Nguyen Le Minh, and Akira Shimazu. 2012. A reranking model for discourse segmentation using subtree features. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL 12, page 160168, USA. Association for Computational Linguistics.

Hongxiao Bai, Hai Zhao, and Junhan Zhao. 2019. Memorizing all for implicit discourse relation recognition. *arXiv preprint arXiv:1908.11317*.

Xue Bai. 2011. Predicting consumer sentiments from online text. *Decision Support Systems*, 50(4):732–742.

Chloé Braud, Ophélie Lacroix, and Anders Søgaard. 2017. Cross-lingual and cross-domain discourse segmentation of entire documents. *arXiv preprint arXiv:1704.04100*.

Harry Bunt and Rashmi Prasad. 2016. Iso dr-core (iso 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings 12th joint ACL-ISO workshop on interoperable semantic annotation (ISA-12)*, pages 45–54.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In *International Workshop on Computational Logic in Multi-Agent Systems*, pages 1–17. Springer.

Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. 2010. Do not feel the trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web*, Shanghai, China.

Lynn Carlson and Daniel Marcu. 2001. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, 54:56.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, pages 85–112. Springer.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Vera Demberg, Merel Scholman, and Fatemeh Asr. 2019. How compatible are our discourse annotation frameworks? insights from mapping rst-dt and pdtb annotations. *Dialogue and Discourse*, 10:87–135.

Takshak Desai, Parag Pravin Dakle, and Dan Moldovan. 2020. Joint learning of syntactic features helps discourse segmentation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1073–1080.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence*, 77(2):321–357.

Judith Eckle-Kohler, Roland Kluge, and Iryna Gurevych. 2015. On the role of discourse markers for discriminating claims and premises in argumentative discourse. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2242.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*.

Nancy L Green. 2018. Towards mining scientific discourse using argumentation schemes. *Argument & Computation*, 9(2):121–135.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A sequential model for discourse segmentation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 315–326. Springer.

Freya Hewett, Roshan Prakash Rane, Nina Harlacher, and Manfred Stede. 2019. The utility of discourse parsing features for predicting argumentation structure. In *Proceedings of the 6th Workshop on Argument Mining*, pages 98–103, Florence, Italy. Association for Computational Linguistics.

Carolin Hofmockel, Anita Fetzer, and Robert M Maier. 2017. Discourse relations: Genre-specific degrees of overtness in argumentative and narrative discourse. *Argument & Computation*, 8(2):131–151.

Alexander Hogenboom, Frederik Hogenboom, Uzay Kaymak, Paul Wouters, and Franciska De Jong. 2010. Mining economic sentiment using argumentation structures. In *International Conference on Conceptual Modeling*, pages 200–209. Springer.

Hsin-Ping Huang and Junyi Jessy Li. 2019. Unsupervised adversarial domain adaptation for implicit discourse relation classification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 686–695, Hong Kong, China. Association for Computational Linguistics.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414.

Tatsuki Kuribayashi, Paul Reisert, Naoya Inoue, and Kentaro Inui. 2018. Towards exploiting argumentative context for argumentative relation identification. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing NLP*, pages 284–287.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1308, Copenhagen, Denmark. Association for Computational Linguistics.

Alex Lascarides and Nicholas Asher. 2008. Segmented discourse representation theory: Dynamic semantics with discourse structure. In *Computing meaning*, pages 87–124. Springer.

Huong LeThanh, Geetha Abeysinghe, and Christian Huyck. 2004. Generating discourse structures for written texts. In *Proceedings of the 20th international conference on Computational Linguistics*, page 329. Association for Computational Linguistics.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Davide Liga. 2019. Argumentative evidences classification and argument scheme detection using tree kernels. In *Proceedings of the 6th Workshop on Argument Mining*, pages 92–97.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2010. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184.

Marco Lippi, Francesca Lagioia, Giuseppe Contissa, Giovanni Sartor, and Paolo Torroni. 2015. Claim detection in judgments of the eu court of justice. In *AI Approaches to the Complexity of Legal Systems*, pages 513–527. Springer.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

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.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. Technical report, University of Pennsylvania Department of Computer and Information Science Technical.

Tobias Mayer, Elena Cabrio, Marco Lippi, Paolo Torroni, and Serena Villata. 2018. Argument mining on clinical trials. In *Computational Models of Argument: Proceedings of COMMA 2018*, pages 137–148.

Jean-Christophe Mensonides, Sébastien Harispe, Jacky Montmain, and Véronique Thireau. 2019. Automatic detection and classification of argument components using multi-task deep neural network. In *3rd International Conference on Natural Language and Speech Processing*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230.

Gaku Morio and Katsuhide Fujita. 2019. Syntactic graph convolution in multi-task learning for identifying and classifying the argument component. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 271–278.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations (ICLR)*.

Philippe Muller, Chloé Braud, and Mathieu Morey. 2019. Tony: Contextual embeddings for accurate multilingual discourse segmentation of full documents. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 115–124. Association for Computational Linguistics.

Akiko Murakami and Rudy Raymond. 2010. Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

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.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):131.

Andreas Peldszus and Manfred Stede. 2016. Rhetorical structure and argumentation structure in monologue text. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 103–112, Berlin, Germany. Association for Computational Linguistics.

Volha Petukhova, Tobias Mayer, Andrei Malchanau, and Harry Bunt. 2017. Virtual debate coach design: assessing multimodal argumentation performance. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 41–50. ACM.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416*.

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 (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. 2005. The Penn Discourse TreeBank as a resource for natural language generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32.

Stephen Purpura, Claire Cardie, and Jesse Simons. 2008. Active learning for e-rulemaking: Public comment categorization. In *Proceedings of the 2008 International Conference on Digital Government Research*, pages 234–243. Digital Government Society of North America.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*, pages 70–77.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. *arXiv preprint arXiv:1704.00217*.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on Empirical Methods in Natural Language Processing*.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE sixth international conference on semantic computing*, pages 30–37. IEEE.

Matthew Rowe and Jonathan Butters. 2009. Assessing trust: contextual accountability. In *Proceedings of the First Workshop on Trust and Privacy on the Social and Semantic Web*, Heraklion, Greece.

Olivia Sanchez Graillet and Philipp Cimiano. 2019. Argumentation schemes for clinical interventions. towards an evidence-aggregation system for medical recommendations. In *HEALTHINFO 2019. The Fourth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing*.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. 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 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 226–234. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

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.

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

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede, Stergos Afantenos, Andreas Peldzsus, Nicholas Asher, and Jérémy Perret. 2016. Parallel discourse annotations on a corpus of short texts. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1051–1058.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Milan Tofiloski, Julian Brooke, and Maite Taboada. 2009. A syntactic and lexical-based discourse segmenter. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 77–80. Association for Computational Linguistics.

Linh Van Ngo, Khoat Than, Thien Huu Nguyen, et al. 2019. Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4201–4207.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Marilyn A Walker, Pranav Anand, Rob Abbott, Jean E Fox Tree, Craig Martell, and Joseph King. 2012. That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Thiemo Wambsganss, Nikolaos Molyndris, and Matthias Söllner. 2020. Unlocking transfer learning in argumentation mining: A domain-independent modelling approach. In *15th International Conference on Wirtschaftsinformatik*.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 17–24.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. *arXiv preprint arXiv:1808.09147*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc.

Appendix: Discourse relation distribution in PDTB 2.0

L1 top-level relations	L2 fine-grained relations	# Instances	
Expansion	***Expansion.Conjunction***	8763	15116
	Expansion.Restatement	3326	
	Expansion.Instantiation	1735	
	Expansion.List	627	
	Expansion.Alternative	531	
	Expansion	118	
	Expansion.Exception	16	
Comparison	***Comparison.Contrast***	5947	7958
	Comparison.Concession	1425	
	Comparison	553	
	Comparison.Pragmatic contrast	21	
	Comparison.Pragmatic concession	12	
Contingency	***Contingency.Cause***	6203	7710
	Contingency.Condition	1359	
	Contingency.Pragmatic cause	78	
	Contingency.Pragmatic condition	68	
	Contingency	2	
Temporal	***Temporal.Asynchronous***	2739	4352
	Temporal.Synchrony	1607	
	Temporal	6	
NoRel	***NoRel***		5464

Table 4: The PDTB top-level (L1) and fine-grained (L2) discourse relations and their distribution in PDTB 2.0 dataset. L2 relations in bold were used for 10-class classification with XLNet.