Generating Discourse Connectives with Pre-trained Language Models: Conditioning on Discourse Relations Helps Reconstruct the PDTB [*]

Symon Jory Stevens-Guille **Aleksandre Maskharashvili** **Xintong Li**
and **Michael White**
Department of Linguistics, The Ohio State University

Abstract

We report results of experiments using BART (Lewis et al., 2019) and the Penn Discourse Tree Bank (Webber et al., 2019) (PDTB) to generate texts with correctly realized discourse relations. We address a question left open by previous research (Yung et al., 2021; Ko and Li, 2020) concerning whether conditioning the model on the intended discourse relation—which corresponds to adding explicit discourse relation information into the input to the model—improves its performance. Our results suggest that including discourse relation information in the input of the model significantly improves the consistency with which it produces a correctly realized discourse relation in the output. We compare our models' performance to known results concerning the discourse structures found in written text and their possible explanations in terms of discourse interpretation strategies hypothesized in the psycholinguistics literature. Our findings suggest that natural language generation models based on current pre-trained Transformers will benefit from infusion with discourse level information if they aim to construct discourses with the intended relations.

1 Introduction

Traditional approaches to discourse have shown the essential importance of discourse (rhetorical) relations in providing coherence to a text (Mann and Thompson, 1987; Lascarides and Asher, 2008; Kehler and Kehler, 2002). While current approaches to natural language generation (NLG) employing pre-trained models have been shown to excel in generating well-formed texts (Kale and Rastogi, 2020, i.a.), their ability to produce coherent texts structured with the help of discourse connectives is understudied (Maskharashvili et al., 2021). The impetus for the present study is the growing body of evidence that neural models, whether trained fresh (Stevens-Guille et al., 2020) or pre-trained (Maskharashvili et al., 2021), benefit from input which includes specific reference to the discourse structure intended to hold in the output text (Balakrishnan et al., 2019). This line of work is novel in the context of current NLG practice, which frequently omits cues to discourse structure in the input. The previous work is purposefully restricted to producing relatively homogeneous texts (museum descriptions and weather predictions). Given the findings of this work on generating limited sets of discourse relations and connectives, it is informative to study the performance of neural models in generating texts structured with the help of a richer set of discourse relations realized by a wide variety of discourse connectives. We study whether having discourse relation information in the input helps neural models to realize the intended discourse relation. These conditions more closely approximate the context in which robust NLG systems would be deployed. We expect our results to provide insight into whether and how to include discourse structure cues in fully-fledged NLG systems.

We report the results of our experiments using BART (Lewis et al., 2019) and the Penn Discourse Tree Bank (Webber et al., 2019) (PDTB) to generate texts with correctly realized discourse relations. We address a question left open by previous research (Yung et al., 2021; Ko and Li, 2020) concerning whether conditioning the model on the intended discourse relation—which corresponds to adding explicit discourse relation information into the input to the model—improves its performance. While we recognize that a positive answer to this question might seem obvious, it has, to date, not been supported with quantitative evidence. We compare our models' performance to baselines in which i) connective choice is determined by the most frequent connective which realizes the intended relation in the corpus, (ii) connec-

[*]E-mail: stevensguille.1@buckeyemail.osu.edu

Proceedings of the SIGdial 2022 Conference, pages 500–515
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022 ©2022 Association for Computational Linguistics

tive choice is determine by the most frequent connective in the corpus irrespective of the intended relation to be expressed, (iii) connective choice is determined by off-the-shelf BART-large mask substitution, and (iv) connective choice is determined by off-the-shelf BERT (Devlin et al., 2019) mask substitution. We propose two types of models by fine-tuning BART on PDTB: models that have discourse relation information in the input (D+ models) and models that do not (D- models). We find that our fine-tuned D+ models substantially outperform fine-tuned D- models, while both kinds of fine-tuned models dramatically beat the baselines. In addition, fine-tuned D+ models produce systematically fewer errors than corresponding D- ones when tested against psycholinguistic observations that certain discourse relations tend to be realized implicitly, while others usually are realized by explicit (overt) discourse connectives. It is important to also point out that our fine-tuned models, unlike previous work and some of our baselines, are not given the position into which the connective should be inserted. This more closely approximates the intended usage of end-to-end neural models, where there is no module in which connectives are slotted into predetermined positions in the output string. We find the models' choices for connective positions to be qualitatively good and focus in the sequel on the connective choices themselves.[1]

2 Background

BART, a transformer-based (Vaswani et al., 2017) language model, is trained on purposefully corrupted data so that the model learns to 'denoise' the corrupted input in the process of reconstructing the original data. Fine-tuning BART on different versions of input and output lets us probe whether the underlying language model needs or benefits from explicit cues to consistently reconstruct the intended discourse connective. The PDTB is one of the few corpora developed to identify discourse dependencies in texts. PDTB provides a well-developed ontology of discourse relations; these discourse relations are used to annotate the Wall Street Journal (WSJ) corpus of the Penn Treebank. We construct versions of the corpus differing in (i) whether the order of the arguments in the output is explicitly encoded in the input, (ii) whether the output is the connective or the connective embedded in the corresponding WSJ text, and (iii) whether a discourse relation is included in the input and how specific it is. The third difference is conceptually the most important one since it corresponds to whether the model is conditioned on discourse relation information.

To determine how well the models realize discourse relations, in addition to standard metrics (i.e., recall and precision), we employ more recent metrics inspired by psycholinguistic (Murray, 1997; Sanders, 2005; Yung et al., 2021) and corpus studies (Asr and Demberg, 2012, 2013; Jin and de Marneffe, 2015a) which allow us to find out the degree to which the models' preferences for realizing different discourse relations correspond to reported human preferences for realizing those relations. In particular, it is argued that while some discourse relations are mostly expressed explicitly, by means of a discourse connective (i.e., overt lexical item or items), other discourse relations tend to be expressed implicitly, i.e., without explicit lexical markers. One of the questions we want to answer is whether providing a discourse relation in the input helps models to learn when to realize a discourse relation explicitly and/or implicitly.

Asr and Demberg (2012, 2013) argue that the PDTB provides ample evidence for psycholinguistic patterns of behaviour. In lieu of directly running human judgement experiments on our model outputs, we test our models' consistency with psycholinguistic results indirectly: we compare the distributions in model outputs to those distributions in the corpus which have been argued to support psycholinguistic theories. We focus on the following two hypotheses:

The Continuity Hypothesis: 'Readers have a bias towards interpreting sentences in a narrative as following one another in a continuous manner . . . additive additive and causal connectives should lead to less processing facilitation than adversative connectives because the former indicate continuity in the discourse whereas the adversatives indicate discontinuity.' — Murray (1997, p.228-229)

The Causality-by-default Hypothesis: 'Because readers aim at building the most informative representation, they start out assuming the relation between two consecutive sentences is a causal relation . . . Subsequently,

[1] In the appendix we provide examples of initial and final connectives which complement the medial connectives used throughout the rest of the paper. We note, however, that our BART-base models prefer producing appropriately positioned initial or medial connectives rather than final connectives.

causally related information will be processed faster, because the reader will only arrive at an additive relation if no causal relation can be established.' — Sanders (2005)

Asr and Demberg (2012) report that in the PDTB, relations that they consider discontinuous and continuous in the sense of Murray (1997) are more likely to be realized explicitly and implicitly, respectively, which is consistent with the continuity hypothesis. Furthermore, they find the proportion of implicit to explicit connectives is highest for causal senses. They conclude this provides support for the hypothesis of causality-by-default.

Asr and Demberg (2013) propose a metric for deriving 'markedness'—how much information about the intended discourse relation is conveyed by a connective or set of connectives—from the PDTB. However, no prior work on the PDTB conditioned models on the discourse relation intended to be communicated. To our knowledge, the following questions, which we report on here, are yet to explored in NLG: (i) whether conditioning on discourse relations improves the prediction of intended discourse connectives; (ii) whether ordering information concerning the arguments to be expressed should be encoded explicitly; and (iii) whether neural models can learn distributions consistent with psycholinguistic results (Sanders, 2005; Murray, 1997) known to be reflected in the training set (Asr and Demberg, 2012).

3 Methods

Our experiments use the BART-base and BART-large implementations of HuggingFace fine-tuned on different versions of the reconstructed PDTB corpus, which we describe in the sequel.[2]The corpus is a modified version of the WSJ texts derived from reconstructing the texts from the string positions provided by the PDTB. Our modifications were intended to make the reconstructions more natural for pre-trained models by using full sentences but without giving away hints to connective location by capitalization or punctuation. An example is provided in Figure 1.[3] The input consists of the set of '<sep>' separated items, while the output is the text the model is trained to produce. The output is either the reconstructed text (marked as full-ouput) or the connective of import in the reconstructed text (marked as conn-only).[4]

We produced in total sixteen different versions of the corpus (see Table 3), twelve of them with discourse relations in the input, which we dub BART$_{D+}$ models, and four without these relations, which we dub BART$_{D-}$ models. Previous work (Asr and Demberg, 2012) found correspondence between the distribution of implicit (respectively explicit) connectives in the WSJ and human behavior reported by Murray (1997); Sanders (2005) concerning which discourse relations are expected (respectively less expected). We produced three versions of the corpus reflecting different levels in the PDTB sense hierarchy as follows:

- Level 1 is the top level (Temporal, Expansion, Comparison, Contingency).
- Level 2 is the set of children of the level 1.
- Full is the set of complete senses, the depth of which is no more than 3.

To determine whether the order of arguments in the PDTB affects the model's choice of connective, we further divided the corpus into versions which included or didn't include explicit encoding of the order of arguments in the output: '12' encodes the case where the first argument precedes the second argument in the reference text, while '21' corresponds to the second argument preceding the fist argument in the reference text. This is in principle useful since the order of arguments in the input need not reflect their order in the output. To control for the influence of generating full texts, we produced versions of the corpus in which the outputs were the discourse connectives without the surrounding WSJ text. Since whether the discourse connective is left implicit or made explicit is something we would like to test every model on, we include no information about whether the connective should be implicit or explicit in the input.

The difference between BART$_{D+}$ and BART$_{D-}$, corresponding to whether the input includes the discourse connective's type,

[2] We find little difference in the performance of BART-base and BART-large and therefore focus on BART-base throughout the paper. Results on matching for BART-large can be found in appendix G.

[3] Due to reconstruction from string positions, the output text is sometimes missing spaces or punctuation from the end of the arguments. The first letter of every argument in the input is in lower case for uniformity. The context arguments can be empty if the string indices from the PDTB correspond to sentences. The string 'none' is inserted into empty contexts. If the PDTB string indices do not correspond to sentences, the context arguments correspond to the sentences in which the PDTB string indices were embedded.

[4] While the PDTB is licensed from the LDC, the scripts for producing our corpora from it plus the metrics and model details will be made freely available on `https://github.com/SymonJoryStevens-Guille/PennGen`.

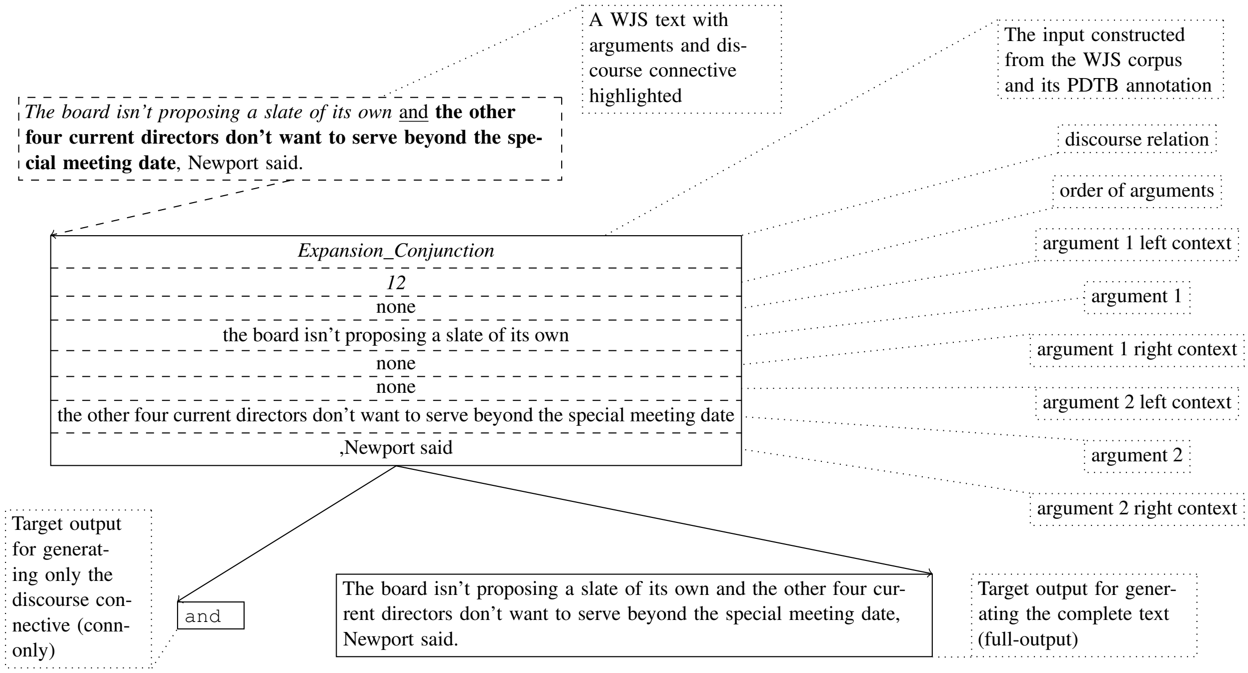
A WJS text with arguments and discourse connective highlighted
The input constructed from the WJS corpus and its PDTB annotation
The board isn't proposing a slate of its own and **the other four current directors don't want to serve beyond the special meeting date**, Newport said.
discourse relation
order of arguments
argument 1 left context
argument 1
argument 1 right context
argument 2 left context
argument 2
argument 2 right context
Expansion_Conjunction
12
none
the board isn't proposing a slate of its own
none
none
the other four current directors don't want to serve beyond the special meeting date
,Newport said
Target output for generating only the discourse connective (conn-only)
and
The board isn't proposing a slate of its own and the other four current directors don't want to serve beyond the special meeting date, Newport said.
Target output for generating the complete text (full-output)

Figure 1: A WSJ text together with its PDTB annotation used in constructing the input to the models and their target outputs. (In the linearized input form, the fields are separated by a <sep> token.)

rounds out the set of distinctions between corpus versions. Details of the corpus split into train/dev/test can be found in Appendix A.

4 Metrics

In order to study whether the models can reconstruct the discourse connectives found in the WSJ, we report model-reference matching. We further consider matching with respect to implicit and explicit discourse connectives.

Implicit and explicit matching, both independently and summed, is our first metric. Second, we consider the proximity of the mismatches between reference and generated connective in terms of the PDTB sense hierarchy. With respect to Figure 1, matching would be producing *and*, either in the embedded sentence in which it occurs (full-output) or on its own (conn-only). Since the type of *and* in this context is Expansion_Conjunction, the mismatch could be by substitution of some other connective from Expansion_Conjunction (=full), from one of the subsenses of Expansion (=level 1), or from some completely different sense (=level 2).

We test the consistency of the models with the continuity and causality-by-default hypotheses by reference to the metrics proposed by (Asr and Demberg, 2012, 2013) to quantify the support for such hypotheses in the PDTB. The usefulness of the distribution of implicit versus explicit connectives is helpfully summed up by Asr and Demberg (2012, pg. 2671): "if readers have a default preference to infer a specific relation in the text, this type of relation should tend to appear without explicit markers." This likewise motivates our use of the metric of markedness, to be discussed below, since markedness quantifies how expected a relation is and, in conjunction with the hypothesis of Uniform Information Density (UID) (Jaeger and Levy, 2006), how likely it is to be explicitly cued versus left to be inferred (Asr and Demberg, 2013).

Asr and Demberg (2012) propose to define implicitness of PDTB senses in terms of the distribution of implicit discourse relations corresponding to the sense in the corpus (# of implicit tokens of senses divided by # of tokens of senses). Following Asr and Demberg (2012, 2013); Jin and de Marneffe (2015b), we focus on the two groups of (sub)types in Table 1, which respectively represent discontinuous and/or noncausal relations and continuous and/or causal relations.[5]

Implicitness and explicitness provide one sort of proxy for continuity and discontinuity in our metrics. We therefore compare the distribution of

[5] We ignore some relations identified by the foregoing authors which don't appear frequently enough in our test set. We report results for the level 1 relations too.

Continuous	Discontinuous
Contingency_Cause	Comparison
Expansion_Instantiation	Temporal_Asynchronous

Table 1: Continuous discourse relation types are shown on the left and discontinuous ones on the right

implicitness predicted by our models to the distribution of implicitness in the test set to determine the fit of the model with respect to the continuity hypothesis.

Following Asr and Demberg (2013); Jin and de Marneffe (2015b), we make use of a metric of markedness, which Asr and Demberg (2013) argue provides a good picture of how likely a given relation is to appear with a connective and to what degree the relation-connective co-occurrence is unique: The higher the markedness, the more likely the relation is to appear with a set of specific connectives. Consequently, it would be more surprising to have that relation cued by a less expected connective or no connective at all—we should then expect both causal and continuous relations to have lower markedness.

Asr and Demberg (2013) report the markedness of level 1 relations, finding the gross cline of Temporal < Contingency < Expansion < Comparison. They argue these results are consistent with the UID, the continuity-by-default hypothesis, and the causality-by-default hypothesis. We consider the degree to which the markedness cline of our model outputs corresponds to the markedness cline of the corpus to provide evidence of whether the model is learning to produce text consistent with the previously mentioned cognitive biases.

Markedness is defined in the equation below, where *npmi* is normalized point-wise mutual information, r belongs to the set of relations R, and c belongs to the set of connectives C minus the null connective.

$$markedness(r) = \sum_{c \in C} p(c|r) \frac{npmi(r;c)+1}{2}$$

Since the markedness metric doesn't provide a direct probability distribution, significance for differences between markedness must be measured by non-parametric methods. For these purposes we use approximate randomization (AR) (Noreen 1989): we randomly re-sample from the two models' union, producing 30K versions of the results and comparing whether and how many such versions improve over the different model predictions in terms of proximity to the reference score (we describe AR at length in Appendix C).

5 Results

With respect to the types of fine-tuning we experimented with, we find BART$_{D+}$ models routinely exceed BART$_{D-}$ models. We show here that BART$_{D+}$ models seem to recover even some of the distributions found by Asr and Demberg (2012) to support psycholinguistic results concerning discourse structure.[6]

In Table 3 we include the results of our baselines. Both models D+ (=80.5%) and D- (=67.2%) significantly improve over the corresponding baseline D+ and D- models. This improvement is further corroborated by comparing BERT and BART-large off-the-shelf to the corresponding BART$_{D+}$ and BART$_{D-}$ models. Both D+ (=79%) and D- (=71.3%) make over a 20% improvement on both BERT and BART-large off-the-shelf. Since the off-the-shelf models were given intended position information in the form of MASK tokens, the result shows that this positional information, at least without fine-tuning, doesn't suffice to predict the intended connective.[7]

Interestingly, with respect to producing matching explicit (respectively implicit) connectives, the models trained to produce full sentence outputs frequently outperform the models trained to produce only discourse connective outputs. This is shown in Table 3, where the difference in scores is most visible when the model is provided with less or no information concerning the intended discourse relation. This suggests there is some benefit to producing the connective in context, where the fidelity of the decoded connective is improved by the preceding and subsequent strings. But this benefit seems to taper off from depth 2 down.

There seems to be a sweet spot in the level of discourse relation type included in the input: there is little improvement between full and level 2 types

[6] The chosen connective need not occur directly between the arguments in the input. Determining which connective is produced by the full-out model is done by iteratively substituting elements of the input found in the output with the empty string. Once this process is complete the remaining strings will include the connective. Strings which are not in the complete set of connectives are removed to eliminate noise. If no connective is found after this process then the model evidently chose to leave the relation implicit.

[7] Note that the MASK position for implicits is uniformly between the rightmost position of arg1 and the leftmost position of arg2. We chose this position for uniformity in light of the absence of implicit connective span annotation in the PDTB.

Type	Comparison	Contingency	Expansion	Temporal
Reference	0.53624	0.21141	0.34245	0.47499
BART$_{D+}$	0.62666	0.21302	0.32155	0.46709
BART$_{D-}$	0.35860	0.19829	0.29256	0.42175

Table 2: Level 1 markedness scores by model

Type	Level	Order	FullOutput	Match
baseline$_{D+}$		-	-	34%
baseline$_{D-}$		-	-	16%
BART-large		-	+	48%
BERT		-	+	52.4%
BART$_{D+}$	full	+	+	79%
BART$_{D+}$	full	+	-	81.5%
BART$_{D+}$	full	-	+	79%
BART$_{D+}$	full	-	-	80.5%
BART$_{D+}$	2	+	+	79.3%
BART$_{D+}$	2	+	-	79.9%
BART$_{D+}$	2	-	+	79%
BART$_{D+}$	2	-	-	79.8%
BART$_{D+}$	1	+	+	76.5%
BART$_{D+}$	1	+	-	66.9%
BART$_{D+}$	1	-	+	75.7%
BART$_{D+}$	1	-	-	65.5%
BART$_{D-}$		+	+	70.5%
BART$_{D-}$		+	-	69.9%
BART$_{D-}$		-	+	71.3%
BART$_{D-}$		-	-	67.2%

Table 3: Model typology with distributions of matched versus reference discourse connectives. BART-large and BERT are baselines used off-the-shelf; baseline$_{D+}$ is the majority baseline conditioned on discourse relations and baseline$_{D-}$ is the majority baseline unconditioned on discourse relations.

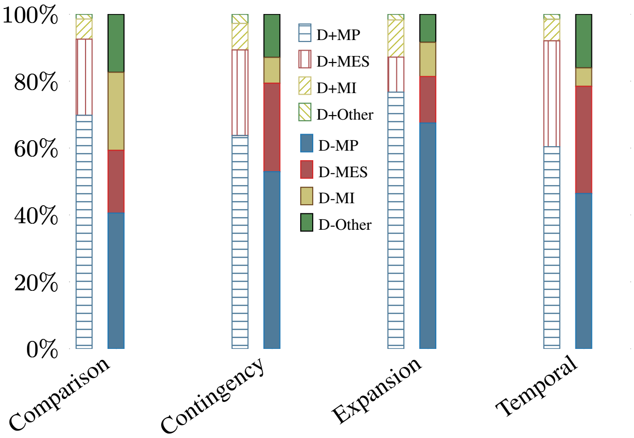

Figure 2: Case of Explicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction); MES (Mismatch with Explicit Sister type); MI (Mismatch with Implicit); and Other (Explicit For Explicit Minus One, Minus Two, and Minus Three level types)

but there is greater improvement between level 2 and level 1—for conn-only the BART$_{D-}$ models even sometimes exceed the level one BART$_{D+}$ models, suggesting the top level type information could hinder connective choice when the connective isn't generated in context.

Despite the boost to connective matching when producing conn-only, the distinction between models which condition on the order of arguments versus those that do not, controlling for other corpus distinctions, is minimal when present. This, too, is visible in Table 3.

The major difference between models with respect to reconstructing the reference connectives is the difference between the BART$_{D+}$ and BART$_{D-}$ models. The BART$_{D+}$ models from the second level to the full level outperform the BART$_{D-}$ models when controlling for the input order, whether the output is full-output or conn-only. In the sequel we report significance results just for the best BART$_{D-}$ model and a corresponding BART$_{D+}$ model: BART$_{D-}$ (-Order,+FullOutput) and BART$_{D+}$ (Depth 3,-Order,+FullOutput). While the level 2 BART$_{D+}$ model ekes out the level 3 model, the difference is uninteresting.

The main distinction in matching between BART$_{D+}$ and BART$_{D-}$ models is due to explicit connectives. Both models perform well with respect to reconstructing implicit connectives, though the differences even here are significant, with the BART$_{D-}$ model even improving over the BART$_{D+}$ model with respect to implicit relations. However, this observation points to a more likely story for BART$_{D-}$'s performance: the BART$_{D-}$ model is less accurate. This is corroborated by it generating an excess of 405 implicits for target explicits compared to BART$_{D+}$'s 275 implicits for target explicits. The overproduction of implicits is further borne out by the differences in F1 shown

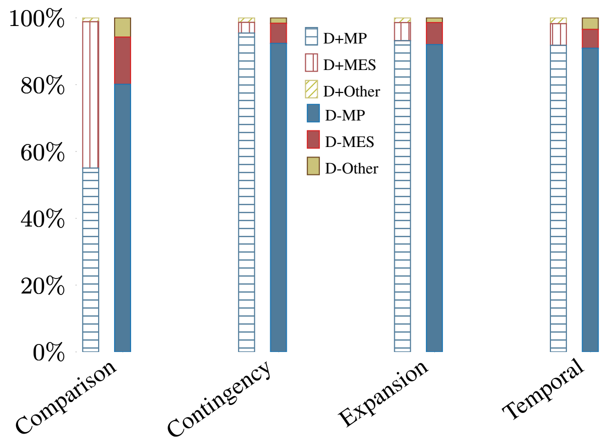

Figure 3: Case of Implicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction), MES (Mismatch with Explicit Sister type), and Other (Explicit For Implicit Minus One, Minus Two, and Minus Three types)

Model	Explicit Match	Implicit Match
BART$_{D+}$	69.8%	89.2%
BART$_{D-}$	54.3%	90.6%

Table 4: Matches for BART$_{D+}$ and BART$_{D-}$.

in Figures 19 and 20 in Appendix D; errors concerning connective choice are exemplified and discussed there too.

Returning to the production of explicit connectives, the improvements of conditioning on discourse structure information are highly significant both with respect to matching per se and with respect to matching explicit connectives. We provide McNemar's test statistics for explicit matches (statistic=157.00, p=0.000), implicit matches (statistic=118.00, p=0.025), and their combination (statistic=313.000, p=0.00) (the tables can be found in Appendix B).

Table 4 shows match results for both implicit and explicit for BART$_{D-}$ and BART$_{D+}$. More fine-grained results are given in Figures 2 and 3. Focusing on the results concerning implicit relations first, the most noticeable difference is with respect to Comparison—the BART$_{D+}$ model produces far fewer matches than the BART$_{D-}$ model. However, within the mismatches here, the BART$_{D+}$ model overwhelmingly produces explicit connectives for reference implicit when the relation can be expressed by both an explicit or implicit connective. In fact the BART$_{D-}$ model makes more severe mismatches on Comparison than the BART$_{D+}$ model, since its productions of explicit connectives for reference implicit are more frequently productions of connectives which simply are not used to express the relation.

BART$_{D+}$: The board isn't proposing a slate of its own **and** the other four current directors don't want to serve beyond the special meeting date, Newport said.

BART$_{D-}$: The board isn't proposing a slate of its own **because** the other four current directors don't want to serve beyond the special meeting date, Newport said.

Figure 4: Full outputs on the input from Figure 1 for BART$_{D+}$ full and BART$_{D-}$ without cues to order. The model's generated connective is bolded.

With respect to producing matching explicit connectives, the BART$_{D+}$ model exceeds the BART$_{D-}$ model on every top level type. When BART$_{D+}$ doesn't produce a matching explicit connective, it is far more likely to produce an explicit connective which expresses the same relation. For each top level type, the severity of the mismatch is less for BART$_{D+}$ than BART$_{D-}$. Without committing to the position that producing an implicit connective for a relation intended to be expressed explicitly is better or worse than producing an explicit connective for a relation intended to be expressed implicitly, we argue that either mismatch is better than producing a connective which is never used to express the intended relation. On this score, the BART$_{D-}$ model is considerably worse—it is consistently more likely to produce a connective not otherwise used to cue the intended discourse relation.

When the metrics are extended to include whether non-matching connectives chosen by the model fit the intended discourse relation, the BART$_{D+}$ model continues to outperform the best BART$_{D-}$ model. When producing non-matching connectives, we find that the chosen connectives of the BART$_{D+}$ model correspond to the intended discourse relations more frequently than those produced by the BART$_{D-}$ models.

We computed markedness scores for outputs of the BART$_{D+}$ and BART$_{D-}$ models. By applying AR significance tests on markedness score–based statistics, we find that the BART$_{D+}$ output on the test data is significantly closer to the reference data than the output of BART$_{D-}$. We show in Table 5 the distribution of markedness

Type	Contingency_Cause	Expansion_Instantiation	Comparison	Temporal_Asynchronous
Reference	0.182	0.071	0.536	0.436
BART$_{D+}$	0.179	0.061	0.626	0.419
BART$_{D-}$	0.170	0.058	0.358	0.365

Table 5: Markedness Scores for continuity and causality-by-default hypotheses.

	BART$_{D-}$		BART$_{D+}$	
Discourse Relation	Nonsister	Implicit	Nonsister	Implicit
Contingency_Cause	4.3%	2.3%	1.9%	2.7%
Expansion_Instantiation	6.1%	1.6%	2.1%	1.6%
Temporal_Asynchronous	**14.2%**	4.7%	2.3%	4.7%
Comparison	**13.9%**	**16.6%**	1.3%	4.2%
Overall	7.8%	6.6%	1.6%	4.5%

Table 6: Error rates in Mispredicted Nonsister and Mispredicted Implicit of the models' performances w.r.t discourse relations associated with the Continuity and Causality-by-default hypotheses. Proportions are with respect to the sum of the items in the test set meant to express the relation type.

for several discourse relations. We should first note that the BART$_{D+}$ markedness scores are consistently closer to the reference scores than the BART$_{D-}$ scores. Second we note that the continuity hypothesis is partially supported, even just considering this limited set of relations: both Contingency_Cause and Expansion_Instantiation are less marked than both Comparison and Temporal_Asynchronous. This is consistent with our hypotheses that continuous and causal relations should be less marked than discontinuous relations. However, like Jin and de Marneffe (2015b); Asr and Demberg (2013), we found less direct support for the causality-by-default hypothesis, since it is not less marked than Expansion_Instantiation. This is at best consistent with a weak form of the hypothesis, since we have not here reported contexts which would discriminate between the conjunction of the causality-by-default and continuity hypotheses versus just the continuity hypothesis. Our conclusions are further reinforced by Table 2, which shows the BART$_{D+}$ model, in particular, is reasonably close to recovering the markedness exemplified in the test set.

One further difference in distribution is predicted by the causality-by-default and continuity hypotheses for those relations that are or are not continuous or causal, exemplified by the relations found in Table 1. Both these hypotheses posit default inferences. Given the apparent reliability of these defaults with respect to psycholinguistic and corpus studies, we'd expect that learning these defaults would reduce the rate of errors for those relations to which the defaults apply. Consequently, we can compare the proportion of errors for continuous and causal relations to that for discontinuous (and non-causal) relations to determine how likely it is the model learned the default. We expect that explicitly representing discourse relations should support the learning of the default since, by hypotheses, the defaults are correlated with specific relations. Table 6 shows the error proportions.

We find that the D+ model shows a lower error proportion with respect to continuous versus discontinuous relations while the D- model shows a higher proportion of such errors, particularly where the relation to be expressed is discontinuous and more marked. We note that the number of D- Nonsister errors on Temporal_Asynchronous, which dwarf the Implicit errors on the same relation, is consistent with the continuity hypothesis in particular since these relations, which are not subject to default inferences, are important to mark explicitly yet are difficult to mark correctly in the absence of an explicit cue to the relation. On Comparison, D- makes a similar number of Nonsister errors, and also makes more than double the number of overall implicit prediction errors. This makes sense if the model recognizes it's important to signal these relations but erroneously treats them as if they were in a default relation where leaving the connective implicit would be more expected. However, we do not have a ready explanation for why Temporal_Asynchronous does not have more implicit D- errors.

The differences in connective choice between models sometimes result in wildly divergent meanings. Figure 4 shows the BART$_{D+}$ full and BART$_{D-}$ outputs for the input in Figure 1. Neither model is conditioned on the order of arguments. The BART$_{D-}$ model's output uses *because* to erroneously communicate that the intentions of the directors cause the intentions of the board, whereas the BART$_{D+}$ model correctly identifies the intentions of the board and the intentions of the directors without suggesting either intention is dependent on the other (generates *and*).

6 Related Work

Ko and Li (2020) reported the limits of GPT-2 (Radford et al., 2019) for generating texts with discourse connectives. Their results concern both fine-tuning and off-the-shelf experiments. For fine-tuning they conditioned the model on prompt-response pairs, testing the subsequently fine-tuned model on the appropriateness of its output responses to input prompts in conversation. For GPT-2 off-the-shelf they fed the first argument and a candidate discourse connective to the model and took the output to be the second argument. They found that GPT-2 more frequently produced connectives consistent with the judgements concerning the discourse relation inferred by human subjects when their agreement on the discourse relation is high. Like Ko and Li, we are interested in discourse relation realization. However, in Ko and Li's approach the position of the discourse connective is explicitly given to the model (it's the mask). Also, Ko and Li's fine-tuned model is restricted to 11 connectives. We condition models on both the discourse relation and the arguments to provide fine-grained control of the discourse without restricting the position of the discourse connective.

Yung et al. (2021) found that GPT-2 diverges from human subjects in its judgements concerning the substitution of connectives which the PDTB does not distinguish by type. This provides presumptive evidence that large pre-trained language models could be limited in reconstructing human judgements concerning the sense of connectives and their substitutability.

7 Conclusion

The main conclusion one can draw from our results is that discourse relation information is essential for consistently generating matching discourse connectives. While large-scale human judgement experiments on our models' predictions are the most obvious next step, the improvement of the BART$_{D+}$ models over the BART$_{D-}$ models with respect to exact matching is encouraging, especially in light of recent results showing that humans don't uniformly accept substitution of discourse connectives which express the same discourse relation (Yung et al., 2021). With respect to whether mere arguments suffice to generate a discourse connective that correctly realizes the discourse relation holding between them, our results indicate that the purely distributional meaning of texts induced by the models under-determines the relation expressed by explicit discourse connectives. Directly conditioning on discourse relations in the input significantly improves the likelihood of the model producing a connective which corresponds to the intended discourse relation. One must note that conditioning on the discourse relation is especially important when the relation is marked, as in these cases the model is apt to predict an incorrect default (causal or continuous) relationship just from the arguments.

As for markedness score–based statistics, we can conclude that the presence of discourse relations in the input helped BART$_{D+}$ to learn the discourse connective distribution patterns of the PDTB. These metrics provide a useful avenue for testing how well generation models recover patterns which hold for a variety of different variables, from discourse relations themselves, to the strength of co-occurrences between discourse relations and the words used to communicate them. To the degree these patterns track cognitive dependencies, they encourage integration of cognitive models of discourse coherence and NLG evaluation.

Acknowledgments

We thank three anonymous reviewers for their feedback. This research was supported by a collaborative open science research agreement between Facebook and The Ohio State University. The last author has been a paid consultant for Facebook while the research was conducted.

References

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684.

Fatemeh Torabi Asr and Vera Demberg. 2013. On the information conveyed by discourse markers. In *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 84–93.

Anusha Balakrishnan, Vera Demberg, Chandra Khatri, Abhinav Rastogi, Donia Scott, Marilyn Walker, and Michael White. 2019. Proceedings of the 1st workshop on discourse structure in neural nlg. In *Proceedings of the 1st Workshop on Discourse Structure in Neural NLG*.

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.

T Jaeger and Roger Levy. 2006. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems*, 19.

Lifeng Jin and Marie-Catherine de Marneffe. 2015a. The overall markedness of discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1114–1119, Lisbon, Portugal. Association for Computational Linguistics.

Lifeng Jin and Marie-Catherine de Marneffe. 2015b. The overall markedness of discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1114–1119, Lisbon, Portugal. Association for Computational Linguistics.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-text pre-training for data-to-text tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102.

Andrew Kehler and Andrew Kehler. 2002. *Coherence, reference, and the theory of grammar*. CSLI publications Stanford, CA.

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

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from gpt-2. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 52–59.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Aleksandre Maskharashvili, Symon Stevens-Guille, Xintong Li, and Michael White. 2021. Neural methodius revisited: Do discourse relations help with pre-trained models too? In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 12–23, Aberdeen, Scotland, UK. Association for Computational Linguistics.

John D Murray. 1997. Connectives and narrative text: The role of continuity. *Memory & Cognition*, 25(2):227–236.

Eric W. Noreen. 1989. *Computer-intensive methods for testing hypotheses : an introduction*. Wiley, New York.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Ted Sanders. 2005. Coherence, causality and cognitive complexity in discourse. In *Proceedings/Actes SEM-05, First International Symposium on the exploration and modelling of meaning*, pages 105–114. University of Toulouse-le-Mirail Toulouse.

Symon Stevens-Guille, Aleksandre Maskharashvili, Amy Isard, Xintong Li, and Michael White. 2020. Neural NLG for methodius: From RST meaning representations to texts. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 306–315, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

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

Frances Yung, Merel Scholman, and Vera Demberg. 2021. A practical perspective on connective generation. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 72–83.

Relation Data set	Compar.	Contig.	Expan.	Temp.
Train	5297	7592	12605	3302
Dev.	1173	1577	2635	784
Test	1195	1616	2563	774

Table 7: Numbers of occurrences of top level relation types in data sets

	m2y	m2n
m1y	1617	663
m1n	157	825

Table 8: m1 = BART$_{D+}$, m2 =BART$_{D-}$.
statistic=157.000, p-value=0.000
Different proportions of total number of entries with explicit matches (reject H0)

A Data sets collection and statistics

The corpus was split into train/dev/test by selecting the first 70 percent of reconstructed lines for training purposes. To prevent content from one split being encountered in another split, any remaining lines in a WSJ article encountered after the line corresponding to the end of train were removed. This technique is used for preventing spillover of content between dev and test, too, which respectively comprise approximately 15 percent of the corpus. Namely, we have 28796 items in the training set, 6169 in the dev set, and 6149 in the test set.

The breakdown of top level relations distributed through the splits is given in Table 7.

We excluded some items from the corpus if the resulting sequences would be too long, if the relations were not extensions of those defined by level 1 in the foregoing, or to prevent possible repetition of content between train, test, and dev splits.

B McNemar's Significance Results

McNemar's significance test results between BART$_{D+}$ and BART$_{D-}$ models are shown in Tables 8, 9, and 10.

	m2y	m2n
m1y	2459	118
m1n	156	153

Table 9: m1 = BART$_{D+}$, m2 =BART$_{D-}$.
statistic=118.000, p-value=0.025
Different proportions of total number of entries with implict matches (reject H0)

	m2y	m2n
m1y	4076	781
m1n	313	978

Table 10: m1 = BART$_{D+}$, m2 =BART$_{D-}$.
statistic=313.000, p-value=0.000
Different proportions of total number of entries with implicit or explicit matches (reject H0)

C Approximate Randomization with respect to Markedness Stats

We want to see whether markedness scores of the outputs models are close to the reference test data. We compute markedness for the test corpus (i.e., gold reference text), t_{mrk}, which is an n-dimensional vector, where n is a number of discourse relation types. We also compute bd^+_{mrk} and bd^-_{mrk} vectors for the outputs of the BART$_{D+}$ and BART$_{D-}$ models on the test corpus, respectively. Then, we calculate the mean square distances between markedness scores of the test corpus and produced ones, i.e., $\delta^+ = MSQ(t_{mrk}, bd^+_{mrk})$ and $\delta^- = MSQ(t_{mrk}, bd^-_{mrk})$. We find that δ^+ < δ^-, which means that the BART$_{D+}$ model output has markedness score at least as close to the test corpus as one of the BART$_{D-}$ model.

To see whether this difference between BART$_{D+}$ and BART$_{D-}$ is significant, we resort to the Stratified Approximated Randomization (AR) approach. We take the list of outputs of BART$_{D+}$ and BART$_{D-}$, call them $d^+_1, \ldots, d^+_k$ and $d^-_1, \ldots, d^-_k$, where k is the size of the test data set. For each i, we randomly assign to c_i either d^+_i or d^+_i, each with 0.5 probability. In this way we obtain a new list $c_1, \ldots, c_k$. We compute the markedness score for $c_1, \ldots, c_k$, call it c_{mrk}. Then, we calculate $\delta^c = MSQ(t_{mrk}, c_{mrk})$. We compare δ^c with δ^+ and δ^-. We do this N (sufficiently large) number of times. If out of N checks, δ^c was less or equal to δ^+ in p-percent of cases, we say that BART$_{D+}$ differs from BART$_{D-}$ with p-significance. (Usually, p is taken to be 5.)

D Discussion of Errors

We consider several examples of errors in D- models and compare them to the same outputs of the D+ model. This discussion is necessarily limited by the length of the outputs. We do not suggest these errors are representative of the models error in general, restricting ourselves to brief quali-

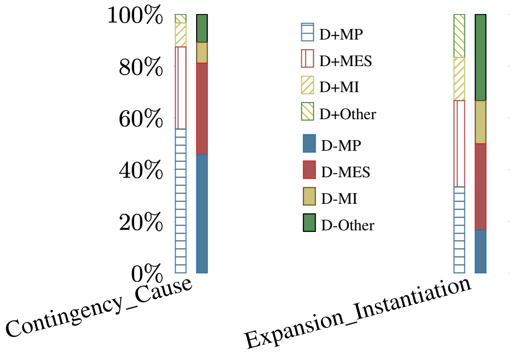

Figure 5: Continuous Explicit Connective Case: Graph for Models D+ and D- showing MP (Match Prediction); MES (Mismatch with Explicit Sister type); MI (Mismatch with Implicit); and Other (Explicit For Explicit Minus One, Minus Two, and Minus Three level types)

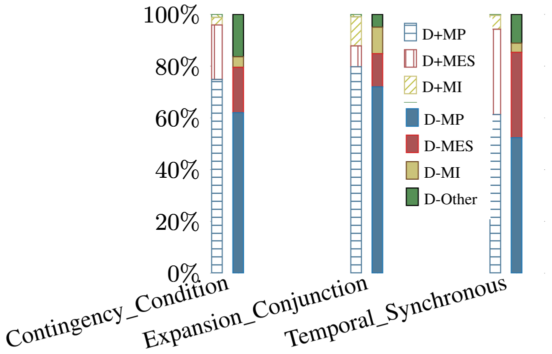

Figure 7: Ambiguous Explicit Connective Case: Graph for Models D+ and D- showing MP (Match Prediction); MES (Mismatch with Explicit Sister type); MI (Mismatch with Implicit); and Other (Explicit For Explicit Minus One, Minus Two, and Minus Three level types)

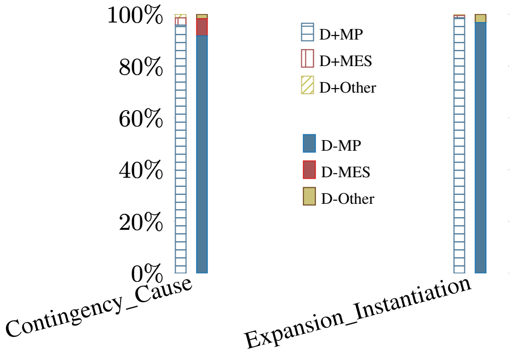

Figure 6: Continuous Case of Implicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction), MES (Mismatch with Explicit Sister type), and Other (Explicit For Implicit Minus One, Minus Two, and Minus Three types)

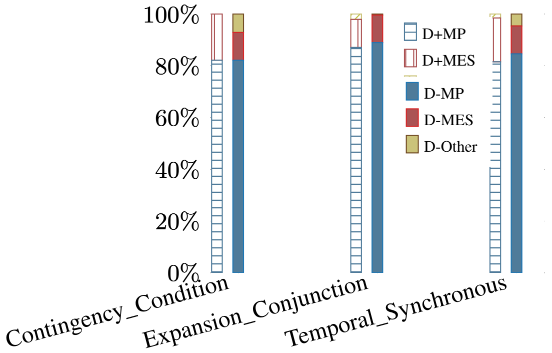

Figure 8: Ambigous Case of Implicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction), MES (Mismatch with Explicit Sister type), and Other (Explicit For Implicit Minus One, Minus Two, and Minus Three types)

tative remarks which complement the quantitative results in the foregoing.

In Figure 9 both models mismatched with the intended temporal_synchronous relation, which is expressed by the connective *while* in the reference text. The D- model's choice produces much more of a hedged judgement of the threat by using *if* than either the reference connective *as long as* or the D+ connective *when*, which seems to require the existence of some time in which the threat is present.

In Figure 10 the D- model mismatched with the intended Comparison_Concession_Arg1-as-denier connective *even if*. The D- model's choice *unless* reverses the intended condition, erroneously suggesting that the banks obtaining financing could prevent British Air from rejecting the proposal described in the text. The D+ model predicts the connective *even if* which matches the reference and communicates the correct dependency between financing and British Air rejecting the proposal described in the text. We note that this is consistent with the results of (Stevens-Guille et al., 2020; Maskharashvili et al., 2021), who found comparison to be quite vexing for LSTM models.

In Figure 11 the D- model mismatched with the intended Expansion_Level-of-detail_Arg2-as-detail connective which is implicit. The D+ model correctly predicts the second sentence to simply provide further comment on the first sentence. Note though that the D- model's connective choice *but* is coherent in the text. This highlights that the two sentences, without the cue to the intended discourse relation, could be understood with respect to a variety of discourse relations.

BART$_{D+}$: Bush assured Roh that the U.S. would stand by its security commitments "*when* there is a threat" from Communist North Korea.
BART$_{D-}$: Bush assured Roh that the U.S. would stand by its security commitments "*if* there is a threat" from Communist North Korea.

Figure 9: Both models mismatch on Temporal_Synchronous, which is expressed by 'while' in the reference text.

BART$_{D+}$: But British Air, which was to have supplied $750 million out of $965 million in equity financing, apparently wasn't involved in the second proposal and could well reject it *even if* banks obtain financing.
BART$_{D-}$: But British Air, which was to have supplied $750 million out of $965 million in equity financing, apparently wasn't involved in the second proposal and could well reject it *unless* banks obtain financing.

Figure 10: D- mismatch on Comparison_Concession_Arg1-as-denier

BART$_{D+}$: The huge drop in UAL stock prompted one takeover stock trader, George KellNER, managing partner of Kellner, DiLeo & Co., to deny publicly rumors that his firm was going out of business. Mr. Kellner said that despite losses on UAL Stock, his firm's health is "excellent."
BART$_{D-}$: The huge drop in UAL stock prompted one takeover stock trader, George Kellner, managing partner of Kellners, DiLeo & Co., to deny publicly rumors that his firm was going out of business. *But* Mr. Kellner said that despite losses on UUAL stock, his firm's health is "excellent."

Figure 11: D- mismatch on Expansion_Level-of-detail_Arg2-as-detail

BART$_{D+}$: The National Cancer Institute also projected that overall U.S. mortality rates from lung cancer should begin to drop in several years *if* cigarette smoking continues to abate.
BART$_{D-}$: The National Cancer Institute also projected that overall U.S. mortality rates from lung cancer should begin to drop in several years *as* cigarette smoking continues to abate.

Figure 12: D- mismatch on Contingency_Condition_Arg2-as-cond

In Figure 12 the D- model mismatched with the intended Contingency_Condition_Arg2-as-cond connective *if*. The D- model's choice of the connective *as* implies that cigarette smoking will continue to abate, while the intended meaning is that the dropping of lung cancer mortality rates in the U.S. depends on cigarette smoking continuing to abate, which abatement, while projected, is not a foregone conclusion.

E Matching Explicit and Implicit Cases of Discontinuous, Continuous, and Ambiguous, Connectives: Figures

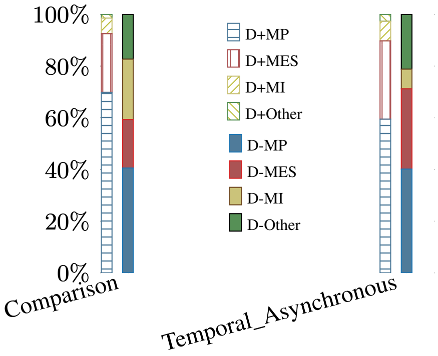

Figure 13: Discontinuous Explicit Connective Case: Graph for Models D+ and D- showing MP (Match Prediction); MES (Mismatch with Explicit Sister type); MI (Mismatch with Implicit); and Other (Explicit For Explicit Minus One, Minus Two, and Minus Three level types)

F Reproducibility Details

We use the pre-trained BART-Large HuggingFace transformer model for our baseline$_{D+}$.

We fine-tuned models, BART$_{D+}$ and $_{D-}$ on BART-Base transformer model. In total, there are 139421184 trainable parameters in this model. The models are fine-tuned using cross entropy loss

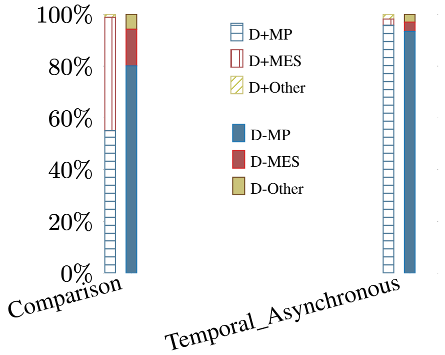

Figure 14: Discontinous Case of Implicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction), MES (Mismatch with Explicit Sister type), and Other (Explicit For Implicit Minus One, Minus Two, and Minus Three types)

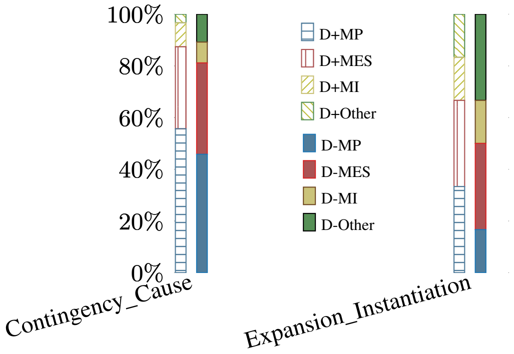

Figure 15: Continuous Explicit Connective Case: Graph for Models D+ and D- showing MP (Match Prediction); MES (Mismatch with Explicit Sister type); MI (Mismatch with Implicit); and Other (Explicit For Explicit Minus One, Minus Two, and Minus Three level types)

without label smoothing. The learning rate is constantly 2×10^{-5} and the batch size is 8 samples. The optimizer is Adam (Kingma and Ba, 2014) where $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and the weight decay is 0. The best checkpoint is selected by validation with patience of 10 training epochs. Computing infrastructure we used is made of NVIDIA V100 GPU and an Intel(R) Xeon(R) Platinum 8268 @ 2.90GHz CPU. Training on average took 15 epochs.

G BART-large selected results

We provide match results for BART-base versions of the full depth D+ and D- models in Table 11.

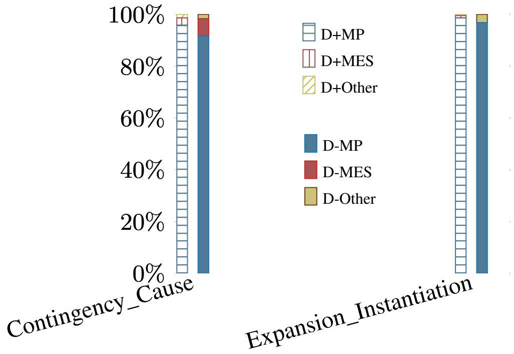

Figure 16: Continuous Case of Implicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction), MES (Mismatch with Explicit Sister type), and Other (Explicit For Implicit Minus One, Minus Two, and Minus Three types)

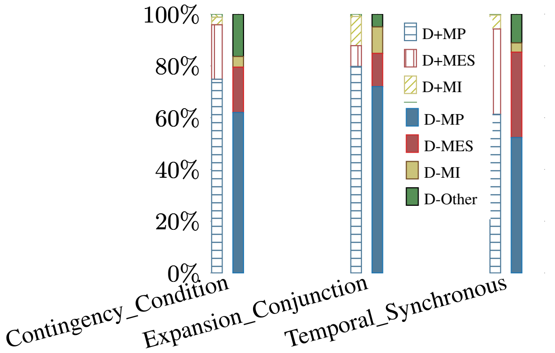

Figure 17: Ambiguous Explicit Connective Case: Graph for Models D+ and D- showing MP (Match Prediction); MES (Mismatch with Explicit Sister type); MI (Mismatch with Implicit); and Other (Explicit For Explicit Minus One, Minus Two, and Minus Three level types)

H Error Rate Examples

Figures 21, 22, and 23 exemplify D- Temporal_Asynchronous Nonsister, Comparison Nonsister, and Comparison Implicit errors respectively.

I Initial and Final Connective Examples

We provide an example of an initial connective generation by D- in Figure 24. A final connective generation by D- is provided in Figure 25, though we note that the reference is here implicit.

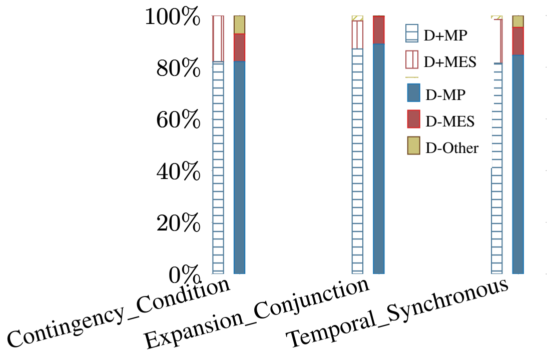

Figure 18: Ambigous Case of Implicit Connectives: Graph for Models D+ and D- showing MP (Match Prediction), MES (Mismatch with Explicit Sister type), and Other (Explicit For Implicit Minus One, Minus Two, and Minus Three types)

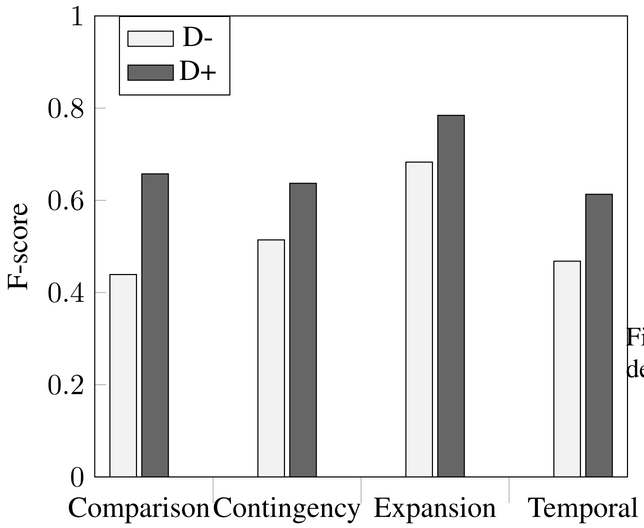

Figure 19: F-score for top level discourse relation types, case of explicit

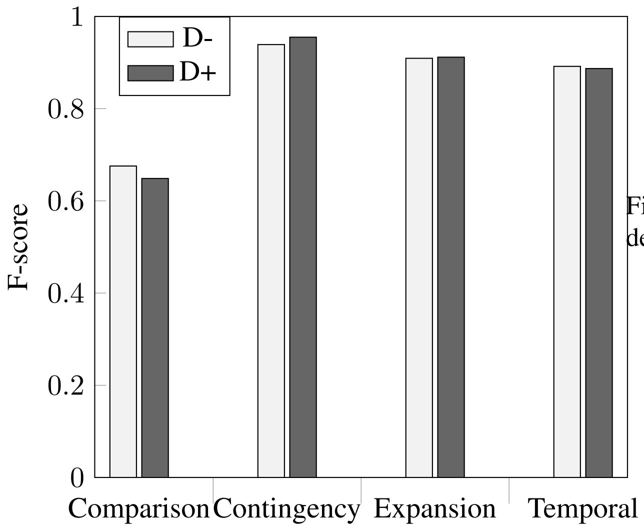

Figure 20: F-score for top level discourse relation types, case of implicit

Type	Depth	Order	FullOutput	Match
BART_{D+}	full	+	+	79.7%
BART_{D+}	full	+	-	82%
BART_{D+}	full	-	+	80.5%
BART_{D+}	full	-	-	74.5%
BART_{D-}		+	+	73.6%
BART_{D-}		+	-	75%
BART_{D-}		-	+	71.5%
BART_{D-}		-	-	74.7%

Table 11: BART-large fine-tuned selected results.

Figure 21: D- Temporal_Asynchronous_Precedence Nonsister Error

REFERENCE: That follows a more subtle decline in the prior six months **after** Manhattan rents had run up rapidly since 1986.

BART_{D-}:That follows a more subtle decline in the prior six months **because** Manhattan rents had run up rapidly since 1986.

Figure 22: D- Comparison_Concession_Arg1-as-denier Nonsister Error

REFERENCE: “There’s quite a bit of value left in the Jaguar shares here *even though* they have run up” lately, says Doug Johnson, a fund manager for Seattle-based Safeco Asset Management.

BART_{D-}: “There’s quite a bit of value left in the Jaguar shares here *and* they have run up” lately, says Doug Johnson, a fund manager for Seattle-based Safeco Asset Management.

Figure 23: D- Comparison_Concession_Arg2-as-denier Error

REFERENCE: But that ghost wasn’t fooled; he knew the RDF was neither rapid nor deployable nor a force — *even though* it cost $8 billion or $10 billion a year.

BART_{D-}: But that ghost wasn’t fooled; he knew the RDF was neither rapid nor deployable nor a force — it cost $8 billion or $10 billion a year.

Figure 24: D- Initial Connective Generation

REFERENCE: But that ghost wasn't fooled; he knew the RDF was neither rapid nor deployable nor a force – *even though* it cost $8 billion or $10 billion a year.
BART$_{D-}$: When Mr. Glass decides to get really fancy, he crosses his hands and hits a resonant bass note with his right hand.

Figure 25: D+ Final Connective Generation

REFERENCE: So far, analysts have said they are looking for $3.30 to $3.35 a share. After today's announcement, that range could increase to $3.35 to $3.40 a share.
BART$_{D-}$:So far, analysts have said they are looking for $3.30 to $3.35 a share. After today's announcement, that range could increase to $4.35 to $2.40 a share **however**.