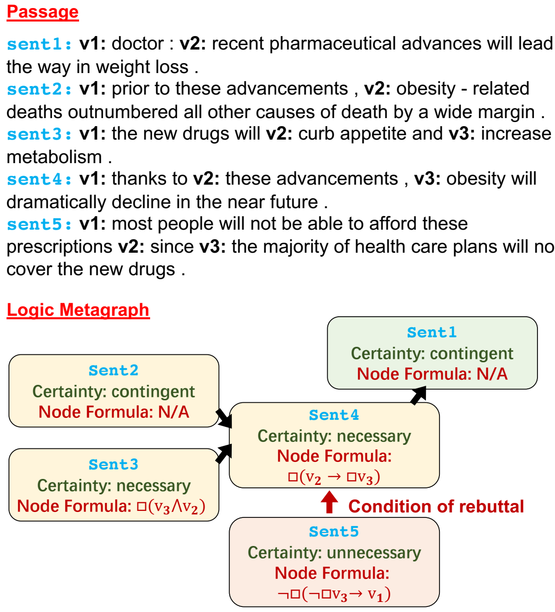
MetaLogic: Logical Reasoning Explanations with Fine-Grained Structure

Yinya Huang[1,2*] **Hongming Zhang**[2†] **Ruixin Hong**[3] **Xiaodan Liang**[1,4†]
Changshui Zhang[3] **Dong Yu**[2]
[1]Shenzhen Campus of Sun Yat-sen University [2]Tencent AI Lab, Seattle
[3]Tsinghua University [4]Pengcheng Laboratory
yinya.huang@hotmail.com, {hongmzhang, dyu}@global.tencent.com,
hrx20@mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn,
xdliang328@gmail.com

Abstract

In this paper, we propose a comprehensive benchmark to investigate models' logical reasoning capabilities in complex real-life scenarios. Current explanation datasets often employ synthetic data with simple reasoning structures. Therefore, it cannot express more complex reasoning processes, such as the rebuttal to a reasoning step and the degree of certainty of the evidence. To this end, we propose a comprehensive logical reasoning explanation form. Based on the multi-hop chain of reasoning, the explanation form includes three main components: (1) The condition of rebuttal that the reasoning node can be challenged; (2) Logical formulae that uncover the internal texture of reasoning nodes; (3) Reasoning strength indicated by degrees of certainty. The fine-grained structure conforms to the real logical reasoning scenario, better fitting the human cognitive process but, simultaneously, is more challenging for the current models. We evaluate the current best models' performance on this new explanation form. The experimental results show that generating reasoning graphs remains a challenging task for current models, even with the help of giant pre-trained language models.



Passage

sent1: **v1:** doctor : **v2:** recent pharmaceutical advances will lead the way in weight loss .
sent2: **v1:** prior to these advancements , **v2:** obesity - related deaths outnumbered all other causes of death by a wide margin .
sent3: **v1:** the new drugs will **v2:** curb appetite and **v3:** increase metabolism .
sent4: **v1:** thanks to **v2:** these advancements , **v3:** obesity will dramatically decline in the near future .
sent5: **v1:** most people will not be able to afford these prescriptions **v2:** since **v3:** the majority of health care plans will not cover the new drugs .

Figure 1: A logical passage and the corresponding logic metagraph in the proposed MetaLogic. Given a logical passage, the goal is to generate the full metagraph including the chain of reasoning with conditions of rebuttal, the node formulae, and the degrees of certainty.

1 Introduction

Being able to generate reasonable explanations is a crucial capability for a reliable reasoning system. Most current works try to ask models to generate reasoning chains as profound explanations. From simple rationales (DeYoung et al., 2020) to more complex multi-step explanations (Inoue et al., 2020; Jhamtani and Clark, 2020; Saha et al., 2021) and deductive chains of reasoning (Clark et al., 2020; Tafjord et al., 2021; Dalvi et al., 2021), previous works attempt to encompass comprehensive information. However, the current explanation design still has limitations for logical reasoning texts in real scenarios. As current explanations lack a fine-grained structure, three remarkable features are not included in current explanations for the sake of real-world logical reasoning: multiple relation types, hierarchical structure, and certainty. As a result, we cannot comprehensively evaluate models' reasoning capabilities in real-life scenarios.

Figure 1 shows examples of the crucial reasoning components that are well studied by previous cognitive science literature (Toulmin, 2003; Garson, 2021) but overlooked by previous work in the machine learning community. First, the inference *rebuttal*. Previous work (Tafjord et al., 2021) mostly only focuses on the inferences of *conjunction* and *entailment* among different statements while ignoring the *rebuttal* ones, which could be crucial in real applications. For example, sent5 counters

[*] This work was done when Y. Huang was an intern at Tencent AI Lab.
[†]X. Liang and H. Zhang are the co-corresponding authors.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4698–4724
December 7-11, 2022 ©2022 Association for Computational Linguistics

`sent4` as a condition of exception and we cannot construct the correct reasoning graph without the *rebuttal* relation. Second, there could exist internal logical relations inside each statement. For example, `sent5` contains two atomic sentences connected by a logical implication relation. Third, real-life statements could have different degrees of *certainty*. For example, "He is hungry" and "He is likely to be hungry" are not identical but relevant because of the certainty. However, most previous work simply treats them completely separately instead of considering their relevance and trying to model the difference (i.e., certainty).

Motivated by previous cognitive science work (i.e., Toulmin Model[1] (Toulmin, 2003) and modal logic theory[2] (Garson, 2021)), we propose a new explanation form, logic metagraphs, to address the aforementioned limitations of previous work. As demonstrated in Figure 1, the logical metagraphs are directed acyclic graphs with meta nodes connected by two types of edges, *support* and *rebut*, representing the inferences between the statements over a logical passage. The meta structure uncovers the chain of reasoning from evidence to the conclusion, along with the challenges from the rebuttal sentences. Each meta node stores information about a logically sound statement formulated as a propositional formula in a standard modal logic S5 system (Hughes et al., 1996), a direct extension of first-order propositional logic with two certainty operators. The formulae have atomic sentences as logical variables that denote events or beliefs, which are modified by three unary operators on their certainty (negation $\neg$, necessity $\Box$, and possibility $\Diamond$) and are joined by three binary operators on their logical relations (implication $\rightarrow$, conjunction $\wedge$, disjunction $\vee$). As a result, the logic metagraphs are comprehensive with multi-hop reasoning paths, inference rebuttal, the internal structure of the statements, and reasoning strength denoted by the degrees of certainty. We collect 1,000 logical passages from the ReClor dataset (Yu et al., 2020) and build the MetaLogic dataset.

Based on our new explanation form, we examine the current best models' ability to understand logical reasoning profoundly. The models need to generate the logic metagraphs given a logical passage. Performances are evaluated by matching scores for the overall structure as well as the three fine-grained components: (1) The inference steps between meta nodes; (2) The per-statement formulae with multiple logical triples; (3) The degrees of certainty. Our evaluation results indicate that generating a comprehensive logical reasoning structure is still challenging for existing giant models.

Our contributions are three-fold:

1. We propose a new explanation form, the logic metagraphs, with a comprehensive logical structure and rich logical information, and the corresponding metagraph generation task.

2. We build a high-quality dataset, MetaLogic, on real-world logical passages.

3. We conduct experiments on three generative models in different frameworks and locate the challenges for current models.

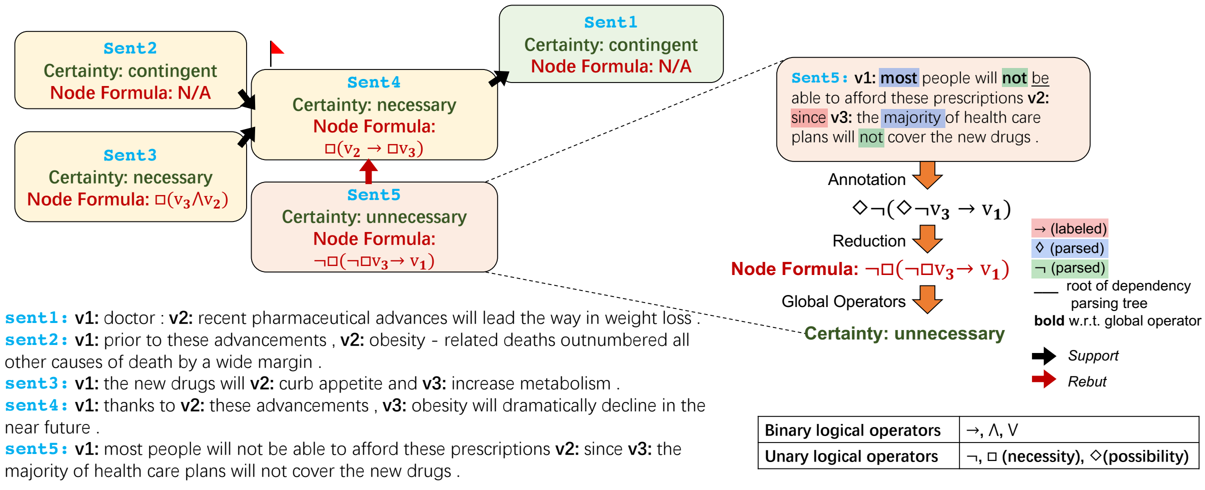
2 Related Works

Explanations Explanation in the context of natural language understanding tasks (e.g., QA) provides interpretability about how models solve the problem. The strategies include asking the models to generate rationales while answering the questions (DeYoung et al., 2020; Inoue et al., 2020), and deriving multi-hop chains of reasoning (Jhamtani and Clark, 2020; Dalvi et al., 2021). The single-sentence rationale provides justification for the question answering but does not uncover the reasoning procedure. While the form of multi-hop chains of reasoning uncovers the reasoning procedure and remedies the simple justification of rationale, it still lacks critical clues about the mechanism within the reasoning steps. Our proposed fine-grained explanation form extends the chain of reasoning by unwrapping the fine-grained texture within each reasoning step. As a result, it allows the reasoning chains to include multiple inference types (e.g., *rebuttal*) and broader reasoning types such as abductive reasoning with the hidden world-knowledge assumption.

Logical Reasoning Machine logical reasoning requires models to conduct hidden symbolic reasoning processes through question answering (Yu et al.,

[1] The Toulmin Model is a canonical theory that helps format and understand arguments. It provides a general pattern to assign logical roles to the sentences in the argument, which clarify the overall logical relations. Especially, the rebuttal components challenge the derivation from existing evidence to the conclusion by providing additional information such as giving a counterexample or proposing an additional condition.

[2] The modal logic theory extends classic first-order propositional logic with two modal operators about certainty and several corresponding rules. This facilitates us to keep the logical variables and relations found in the text and, at the same time, introduce degrees of certainty to the graph.



sent1: **v1:** doctor : **v2:** recent pharmaceutical advances will lead the way in weight loss .
sent2: **v1:** prior to these advancements , **v2:** obesity - related deaths outnumbered all other causes of death by a wide margin .
sent3: **v1:** the new drugs will **v2:** curb appetite and **v3:** increase metabolism .
sent4: **v1:** thanks to **v2:** these advancements , **v3:** obesity will dramatically decline in the near future .
sent5: **v1:** most people will not be able to afford these prescriptions **v2:** since **v3:** the majority of health care plans will not cover the new drugs .

Binary logical operators	→, ∧, ∨
Unary logical operators	¬, □ (necessity), ◇(possibility)

Figure 2: The overall logical reasoning explanation task is defined as follows. Given a passage, a model reconstructs the fine-grained logical structure with the meta *support* or *rebut* relations, the inner node formulae, and degrees of certainty for each node. Given a logical statement, the formula is constructed from the labeled logical triples with the parsed unary operators, which can then be reduced to canonical forms. The certainty label should follow the global operators.

2020; Liu et al., 2020; Cui et al., 2020), or explicitly perform symbolic reasoning via natural language (Clark et al., 2020; Tafjord et al., 2021; Dalvi et al., 2021). The QA-based reasoning data is mostly collected from real-life scenarios without corresponding structural information. To perform reasoning, symbolic modules (Huang et al., 2021; Ouyang et al., 2021) or learning strategies (Wang et al., 2022) are designed to approximate the reasoning structure. On the other hand, explicitly generating chains of reasoning can better uncover models' reasoning processes. However, recent work mostly focuses on deductive reasoning, where models with iterative strategy (Tafjord et al., 2021) or reasoning modules (Hong et al., 2022) show superior performances. To encourage more advanced reasoning capabilities, we propose a comprehensive reasoning structure with fine-grained factors.

Argumentation / Discourse Structures Previous works (Lawrence and Reed, 2019; Li et al., 2022) such as argumentation mining (Stab and Gurevych, 2014b,a, 2017) or discourse parsing (Carlson et al., 2001; Webber et al., 2019) study document structure prediction. Given a passage, a model is required to predict the argument components or the discourse relations between them. Instead of identifying the rhetorical structure of a passage, the proposed logic metagraphs aim at simulating the logical reasoning process, where the model needs to select the relevant knowledge out of a pool to finish the reasoning. Besides, unlike directly considering a sentence or a text span as a reasoning node, MetaLogic explores a schema with finer granularity. Each reasoning node is further decomposed into logical variables with relations and modal operators so that the inner structure as well as the certainty are considered.

3 Task Definition

Overall Generation Task The desideratum is that a model reconstructs the fine-grained logic explanation for a given passage, which uncovers the model's understanding of the logic between the lines. The logic explanation is formatted as logic metagraphs with *support* or *rebut* inference steps, per-node logical formulae, and degrees of certainty, as demonstrated in Figure 2.

The input for the models is a passage with multiple statements $(S^{(0)}, S^{(1)}, ..., S^{(N)})$ and atomic sentences $p_*^{(n)} \subseteq S^{(n)}$, according to which they generate the logic metagraph. The logic metagraph has three main components: (1) The meta structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} = \mathcal{E}_S \bigcup \mathcal{E}_R$, and $\mathcal{E}_S$ and $\mathcal{E}_R$ are the two meta edge types, *support* and *rebut*, respectively, between the meta nodes $u^{(n)} \in \mathcal{V}, n \leq N$. (2) The set of node formulae $\mathcal{F}$, where $u^{(n)} := f_n \in \mathcal{F}$. Each formula is joined by logical triples. $f_n = \bigcap \mathsf{r}(\mathsf{m}(p_i^{(n)}), \mathsf{m}(p_j^{(n)}))$, where $i \neq j$, $\mathsf{r} \in \{\rightarrow, \wedge, \vee\}$, and m is a combination in $\{\neg, \Box, \Diamond\}$. (3) The set of degrees of certainty $\mathcal{C}$, defined by the combination format of $\{\neg, \Box, \Diamond\}$.

Senses				
	Classic	Morality	Tense	Belief
$\Box p$	The proposition p is *necessary*.	p is morally *obligatory*.	It will *always* be the case that p.	Things a person *knows to be true*.
$\Diamond p$	The proposition p is *possible*.	p is morally *permissible*.	It will *sometimes* be the case that p.	Things that *may be true* as far as a person knows.
Definitions				
$\Box p := \neg\Diamond\neg p$ $\Diamond p := \neg\Box\neg p$		It is necessary that p. := It is not possible that not-p. It is possible that p. := It is not necessary that not-p.		
Reduction Rules				
$\Box\neg p = \neg\Diamond p$, $\Diamond\neg p = \neg\Box p$, $\Box\Box p = \Box p$, $\Diamond\Diamond p = \Diamond p$, $\Box\Diamond p = \Diamond p$, $\Diamond\Box p = \Box p$.				
Degrees of Certainty				
$\Box$:= 4 (necessary), $\Diamond$:= 3 (possible), N/A := 2 (contingent), $\neg\Box$:= 1 (unnecessary), $\neg\Diamond$:= 0 (impossible)				

Table 1: Senses, mutual definitions, reduction rules, and degrees of certainty of modal logic operators.

4 The Logic Metagraph

In this section, we introduce the proposed logic metagraph in details.[3]

4.1 Meta Node and Edge

Each meta node corresponds to a logically sound statement (e.g., *premise*, or *conclusion*). The meta edges are either *support* or *rebut*, relating to a single step of inference. The *support* edges join the meta nodes to form a chain of reasoning to the conclusion, whereas the *rebut* edges indicate challenges from the condition of rebuttal to one of the meta nodes in the chain, which are evidence or claims about exceptional conditions. Each inference step allows multiple premises.

4.2 Internal Structure of Meta Node

The internal structure of a statement is formulated as a propositional logic formula. The logical variables denote the atomic sentences in the statement that corresponds to separate events or beliefs. The logical relations between such events or beliefs are denoted by binary propositional operators. There are three logical relations: logical implication, conjunction, and disjunction ($\rightarrow$, $\wedge$, $\vee$). Multiple such logical triples are joined by conjunctions ($\wedge$). Furthermore, each logical variable and the overall formula are modified by negation ($\neg$) and modal ($\Box$ and $\Diamond$) operators, representing the degrees of certainty of each atomic sentence as well as the whole statement, respectively. A more detailed introduction can be found in Section 4.3.

4.3 Certainty with Modal Operators

Modal logic (Garson, 2021) is an extension of first-order propositional logic with two modal operators, necessity ($\Box$) and possibility ($\Diamond$). They are unary operators, and Table 1 presents examples of their senses in natural language (Hughes et al., 1996). For example, $\Box p$ denotes that the proposition p is necessary, while $\Diamond p$ means p is possible, in the classic definition. In another sense of tense, $\Box p$ represents that the evidence p is true at all times, whereas $\Diamond P$ represents that p is only true sometimes. In general, the modal operators indicate certainty information of the propositions.

The two modal operators can define each other with the negation operator ($\neg$). Multiple reduction rules are defined. As a result, any complex formulae composed of modal operators could be reduced to one of the five degree-of-certainty forms listed in Table 1, which is also known as the classic S5 system (Hughes et al., 1996) and makes the logic metagraph defined in a complete set.

5 MetaLogic

In this section, we introduce the construction details of the MetaLogic dataset. Since the logic metagraphs have fine-grained structures with multiple evaluation dimensions, which are all dispensable and supplement each other, we design a rigorous annotation process for the construction.

5.1 Preparation

Source Data We use ReClor (Yu et al., 2020) as the source data, where the multiple-choice questions are collected from GMAT and LSAT. As a

[3] An example is shown on the left side of Figure 2.

pilot study on logical reasoning explanation, we start with the standard text questions so that the explanation form can benefit from precise and comprehensive logical information. Each question contains a logical passage, a question, and multiple answer options. The original dataset contains 17 reasoning types, which can be mainly categorized into two folds: complete reasoning composed of the logical passage and the option (e.g., the types Necessary Assumptions, Sufficient Assumptions, Strengthen, Weaken); flawed in-context reasoning structure (e.g., the types Technique, Identify a Flaw, or Dispute). As we aim to study models' understanding of the complete reasoning process over the whole passage, we consider data from the first category, from which we randomly choose 1,000 samples. Examples of the selected questions can be found in Appendix A.

Data Preprocessing We first filter out incoherent options from the questions for logical structure coherence. For ordinary questions, the incoherent options are the distracting ones. Conversely, for the inverse questions with "EXCEPT", we randomly select one of the distracting options and remove the others. We further split the passage into sentences as the initial meta nodes and per meta node sentence into clauses as the initial logical variables. This follows the convention of applying linguistic-based segments as reasoning components in related studies (Dalvi et al., 2021; Huang et al., 2021; Wang et al., 2022; Xu et al., 2022). Besides, considering the label hierarchy that the logical variables are conditioned on the meta nodes, the initial segments help build the desired metagraph sketch. Moreover, the initial delimitation is trivial with punctuation marks and provides the least machine guidance to the annotators, who are free to modify the segments on their understanding of reasoning units, which will be demonstrated in Section 5.2. From the experts' view, 27 of 30 randomly sampled annotated graphs are of high quality, which indicates the high reliability of starting with the initial segments.

As a result, the text presented to the annotators contains the original text with the passage, the question, and the coherent option, along with a list of delimited sentences.

5.2 Annotation

As all annotation tasks require a global understanding of the overall passage, we recruit the same annotator to finish all tasks in the same passage. The annotation procedure has four steps. (1) Read through the text and have a rough idea about the logical role of each initial meta node (e.g., being a *conclusion* or *rebuttal*). If an initial meta node does not provide complete evidence, then the annotator needs to merge it with another node to form complete evidence. (2) Annotate the inference types between the meta nodes. After this stage, we obtain the chain of reasoning and the rebuttal steps. (3) For each meta node, annotate the logical variables by refining the span boundaries of the given initial logical variables. (4) Annotate the logical binary operator between the logical variables. The annotation platform is demonstrated in Appendix D.

	Meta Structure		Meta Node	
	M-Node	M-Edge	L-Variable	L-Relation
κ	57.80†	42.82†	65.46‡	56.81†

Table 2: IAA with Cohen's Kappa coefficients. M-Node: meta node, M-Edge: meta edge, L-Variable: logical variable, L-Relation: logical relation. ‡ indicates very high agreement with κ over 60%. † indicates high agreement with κ between 40% and 60%.

We recruit annotators from crowd-sourcing platforms. We first train annotators with a carefully designed annotation guideline[4] and require them to pass an exam before the annotation to guarantee the annotation quality. For each passage, we invite two annotators[5]. On average, we pay $2.2 for each logical passage.

For unary logical operators ($\neg$, $\Box$, $\Diamond$), as discussed by (Toulmin, 2003), there exist conventional clue words for the negation and modality. Following that, we leverage such in-context clue words for the annotation. Given a set of conventional indicators (demonstrated in Table 13 in Appendix C), we parse each meta node sentence into a dependency parsing tree, then detect those words within 3-hops to the root node, and assign the corresponding operators to the formula. The consecutive unary operators are ordered by the distance from the indicators to the parsing root node. This results in the global unary operators. For local unary operators of the logical variable spans, we parse the spans and evaluate the indicator-root distance. The repeatedly detected indicators are reduced, as a result, the operators are kept by the global formula and removed

[4] Details are shown in Appendix B.

[5] For the inconsistent annotation, we invite a third annotator to make the judgement.

	Component 1				Component 2		Component 3			Overall
	Node		Step		Formula		Certainty			
	F1	All	F1	All	F1	All	Acc	All	F1*	All
Once (large)	91.0	55.7	45.3	15.0	58.8	57.8	79.2	40.8	24.3	4.4
Multitask (large)	92.8	62.5	52.3	20.7	77.2	75.8	82.8	52.0	41.5	8.9
MetGen (large)	94.5	68.7	56.3	22.3	78.3	76.7	84.0	56.3	50.3	11.2
Once (11b)	92.5	59.0	52.4	23.5	71.0	69.6	84.1	55.0	37.2	11.7
Multitask (11b)	93.9	66.5	58.3	28.0	77.8	75.7	82.8	54.5	45.6	13.2
MetGen (11b)	94.4	68.5	61.5	28.0	80.8	78.8	85.7	60.0	55.7	15.4

Table 3: Evaluation results of generative models. All: AllCorrect. *: macro-F1.

before the local variable. To evaluate the labels, the annotators check 391 unary operators from 200 randomly sampled passages. As a result, 92.6% of them are consistent with human cognition, which indicates that the operators are valid and consistent.

5.3 Inter-Annotator Agreement

We evaluate the inter-annotator agreement in multiple dimensions with Cohen's Kappa Coefficient (Cohen, 1960).

Meta Node The IAA of meta nodes reflects one's understanding of the logical role of each statement. We evaluate the annotators' agreement of each meta node of being one of the five characters: conclusion, rebuttal, beginning of the chain, an intermediate conclusion in the chain, and irrelevant node.

Meta Edge We consider the exhaustive meta node pairs except the reflexive ones. Consequently, the agreement is calculated on the adjacency matrix of the meta edges regarding the three labels: *support*, *rebut*, and without-an-edge. The diagonal elements in the matrix are excluded.

Logical Variable As the logical variables are text spans, the annotators vote for each token for being in a logical variable or not. The agreement is average over the per-token agreement.

Logical Relation Similar to meta edge, we consider the exhaustive logical variable pairs except for the reflexive ones. Considering the logical variables as vertices, the agreement is calculated on the adjacency matrix regarding the four labels: logical implication, logical conjunction, logical disjunction, and without-a-relation. The diagonal elements are regarded.

We present the results in Table 2. The agreement is consistently high, which indicates the high quality of MetaLogic. Moreover, the high IAA also indicates that humans could easily solve the logical reasoning explanation and provide consistent logical reasoning graphs.

Passage	Graph	Node	Form	Reb	Multi-step	Multi-premise
1,000	1,000	3,609	1,500	416	435	400

Avg. Node	Avg. Form	Avg. Var	Avg. Binary	Avg. Unary Global	Local
3.61	1.5	2.63	1.89	2.19	1.31

Table 4: Label statistics. The first row indicates the overall numbers of passages (passage), logic metagraphs (graph), meta nodes (node), formulae (form), number of metagraphs that have *rebuttal* steps (reb), have multi-step chains (multi-step), and have multiple premises (multi-premise). Note that reb, multi-step, and multi-premise have small intersections and $n_{\text{reb}\cap\text{step}\cap\text{premise}} = 36$. The second row shows per-passage average number of meta nodes, formulae, logical variables (var), binary operators (binary), global and local unary operators (global, local).

5.4 Dataset Statistics

The final annotated MetaLogic contains 1,000 logic metagraphs with over 3,609 meta nodes and 1,500 formulae. In the MetaLogic, 416 out of 1,000 logic metagraphs have rebuttal steps. Around 40% of metagraphs have multi-hop reasoning chains. On average, each logic metagraph has more than three meta nodes, and about 40% are mapped to formulae. Moreover, each metagraph has an average of 2.19 global operators. More statistics can be found in Table 4. We randomly split the data with 60% training, 20% development, and 20% testing.

6 Experiment

We evaluate the performance of the following explanation generative models on MetaLogic: (1) All-at-Once (**Once**) T5 (Raffel et al., 2020), which performs sequence-to-sequence generation via generating the whole metagraph in a linearized sequence given the overall passages with the sentence and variable denotations; (2) **Multitask** T5 (Raffel

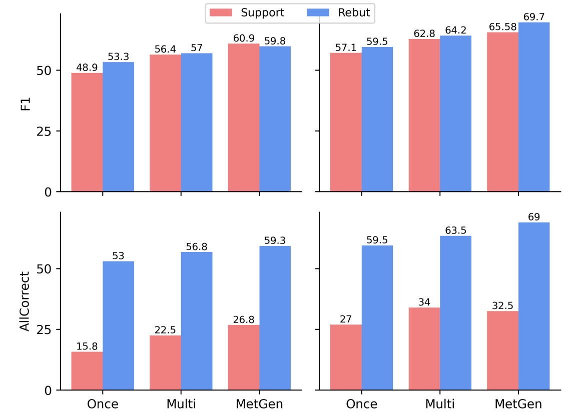


Figure 3: Performances on different inference types: *support* and *rebut*. Left column: models with the T5-large backbone. Right column: models with the T5-11b backbone.

et al., 2020), which complete the whole generation task with a combination of three sub-tasks: meta structure generation, formula generation, and certainty prediction; (3) **MetGen** (Hong et al., 2022), which is a module-based framework for structured explanation generation. It further introduces a reasoning controller and two modules for meta structure generation. To the best of our knowledge, MetGen is the current state-of-the-art explanation generative model. Further model details are in Appendix E. Following Dalvi et al. (2021), we report the F1 and AllCorrect scores for each dimension and the overall AllCorrect score. For certainty, we report the accuracy of a five-label classification and an extra macro-F1 due to the unbalance of the degree labels. The overall AllCorrect is the strictest metric since any difference in the predicted metagraph will make the prediction a wrong one. Details can be found in Appendix F.

6.1 Implementation Details

We fine-tune Once (large), Multitask (large), and MetGen (large) with a batch size of 32 for 300 epochs on 1 Tesla V100 GPU, and fine-tune Once (11b), Multitask (11b), MetGen (11b) with a batch size of 4 for 300 epochs on 8 Tesla V100 GPUs. The learning rate is 1e-5 for all models. The model parameters are optimized by Adafactor (Shazeer and Stern, 2018). The models are evaluated per 10 epochs on the development set, and the best checkpoints are saved for test set evaluation.

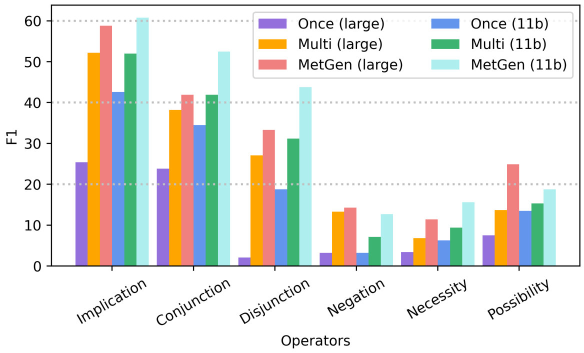


Figure 4: Performances on each logical operator.

6.2 Main Results

From the results shown in Table 3, we can find out that generating the comprehensive logic graph is still a very challenging task, even for current giant models as all models achieve low AllCorrect performance. Specifically, we can make the following observations:

1. From the experiments in certainty prediction, we can see that all models are struggling, which shows that knowing certainty is still not a trivial task for current models given that there are explicit indicators in context.

2. We notice that using larger pre-trained models (e.g., T5-11B) can help improve the performance of all models, this indicates that big models can help better model the statement semantics such that they can better identify and link statements.

3. We also notice that the module-based method MetGen can outperform the Once and Multitask method, which indicates that iteratively generating the explanation graph with basic modules is a more reliable logical reasoning framework.

4. From the experiments on Component 1, we can see that the models could obtain high node scores and mediocre step scores, but the step AllCorrect results are inferior. This indicates that with the help of giant pre-trained LMs, current models could effectively learn to identify the nodes, but they may not know the true logical reasoning because they cannot precisely predict the inference types among these nodes.

5. The models achieve around or over 60% of F1 and AllCorrect scores in predicting the formula, showing their awareness of the inner logical

structure. This makes sense because the majority of the inner structure is triggered by connective words such as “so.”

In the rest of this section, we present a more detailed inspection from different perspectives.

6.3 Performances on Metagraph Parts

We further inspect models’ performance on detailed components. The evaluation results on separate inference types (*support* and *rebut*) are demonstrated in Figure 3. Overall, identifying *rebut* is easier than *support*, according to the exact match scores F1 and AllCorrect. This makes sense because most *rebut* nodes could contain informative keywords such as “however” but the majority of nodes in *support* edges do not. Besides, the average F1 scores per operator are shown in Figure 4. From the results, we can see that the trend of models’ performance are generally consistent on different operators, which indicates that different operators may have different intrinsic difficulty.

6.4 Data Scale for Logical Inference

To investigate how well current models can learn to generate the reasoning graphs, We use different ratios of training data to train the models and present the results in Figure 5. Overall, the model performances show a rapid increase within 20% of training data, then a flat and steady increase and do not reach a platform, indicating that the models can still benefit from more structural reasoning data. Among the models, MetGen has the most significant growth trend and performs data efficiently with small data, showing the advantages of the module-based learning framework in symbolic reasoning. Interestingly, we find out that the performance of multitask T5 decreases after seeing half of the training data. A possible explanation is that the decomposed logical structure as independent sub-tasks prevents the models from a holistic understanding of the logical passages. Besides that, the flat increasing rate after seeing 20% of the training data also suggests that blindly increasing the training data scale may not be the most efficient way of teaching models to conduct such a complex reasoning task.

6.5 Error Analysis

To better understand current models’ errors, we randomly sample 50 instances from the development set and collect the predictions from the All-at-Once (T5-11b) model. We manually evaluate the predictions and categorize 4 to 5 error types for each component, as shown here in the Table 5.

Graph (G1-G5)	Formula (F1-F4)	Certainty (C1-C4)
13/13/11/10/12	32/6/8/4	5/30/10/5

Table 5: Error type statistics. We randomly select 50 test set predictions and group the error types by components. G1: Incorrect inference type. G2: Incorrect rebuttal. G3: Incorrect conclusion. G4: Incorrect inference step. G5: Other structural mismatch. F1: Incorrect logical variable. F2: Incorrect unary operator. F3: Incorrect binary operator. F4: Incorrect implication direction. C1: Incorrect polarity. C2: Other polarities to contingent. C3: Contingent is predicted as other polarities. C4: Unresolved degree of certainty.

Specifically, the meta graph structure mainly has five error types: (G1) Incorrect inference type: the model predicts the correct structure, but over one of the inference steps has the inverse type (i.e., predicted *support* but should be *rebut* or vice versa); (G2) Incorrect rebuttal: missing or incorrectly predict a rebuttal step; (G3) Incorrect conclusion: mismatched conclusion node at the end of the reasoning chain; (G4) Incorrect inference step: missing or predicting redundant inference step; (G5) Other structural mismatches: Including different chain branches and so forth. The four error types in formulae are (F1) Incorrect logical variable: missing or predicting redundant logical variable, or predicting a wrong variable; (F2) Incorrect unary operator: The variables are correct but are bound by incorrect unary operators; (F3) Incorrect binary operator: The variables and unary operators are correct, but predict incorrect binary operator. (F4) Incorrect implication direction: The variables, unary and binary operator types are correct, but the implication operator has an inverse direction. The four error types in certainty: (C1) Incorrect polarity: Predicting the certainty in an opposite polarity; (C2) Other polarities to contingent; (C3) Contingent is predicted as other polarities; (C4) Unresolved degree of certainty.

From the results we can see that, the model tends to predict the operator quite well (F2/F3), but not the variable (F1), which suggests that even though current deep models can identify the correct relations with some trigger words (e.g., “so that”), they may not fully understand it because they cannot find the correct variable span in the context. Be-

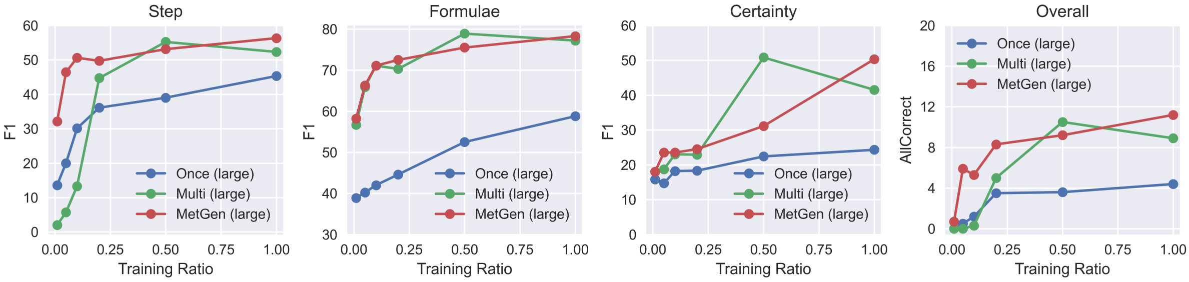


Figure 5: Results on different ratios (0.01, 0.05, 0.1, 0.2, 0.5, 1.0) of MetaLogic training data.

sides that, we also notice that the model tends to predict the wrong polarities, which is typically irrelevant towards the conclusion, as the important certainty feature. This suggests that the model may learn to answer questions with the wrong reason (i.e., short path (Lovering et al., 2021)), which further demonstrates the importance of our task for constructing a reliable and trustworthy reasoning system.

7 Conclusion

This paper extends the boundary of current research on logical graph generation for reliable reasoning systems. Specifically, we carefully design a complete logic explanation form following previous research on cognitive science. Accordingly, we built MetaLogic with a comprehensive annotation task design and quality examination. We also evaluate several recent models and show that the performance of current models is still unsatisfactory, even with giant pre-trained language models. We hope that this paper could motivate more future works on reliable reasoning systems that could generate the correct logical graphs to support their reasoning. The MetaLogic data and implementation code are available at `https://github.com/tencent-ailab/MetaLogic`.

8 Limitation

The major limitation of MetaLogic is that we cannot annotate a large enough dataset for data-driven methods. However, considering that humans could learn to conduct logical reasoning after seeing a few examples, we argue that it is meaningful to investigate whether machines can learn the same level of reasoning capability with limited data.

9 Ethical Considerations

During the annotation process, we follow the minimum payment requirement of the united states. No personal or confidential information is collected. Hence, to the best of our knowledge, there is no ethical concern.

Acknowledgements

We appreciate the anonymous reviewers for their insightful comments. We thank Dr. Su Wu for reviewing the annotation manual, thank Dr. Xingchi Su for double-checking the operator reduction, and thank Jianheng Tang, Zhicheng Yang, and Xinran Zhao for their constructive advice for the manuscript. This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700, National Natural Science Foundation of China (NSFC) under Grant No.61976233, Guangdong Province Basic and Applied Basic Research (Regional Joint Fund-Key) Grant No.2019B1515120039, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), Shenzhen Fundamental Research Program (Project No. JCYJ20190807154211365) and CAAI-Huawei MindSpore Open Fund. We thank MindSpore for the partial support of this work, which is a new deep learning computing framework[6].

References

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the SIGDIAL 2001 Workshop, The 2nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, Saturday, September 1, 2001 to Sunday, September 2, 2001, Aalborg, Denmark*. The Association for Computer Linguistics.

[6] https://www.mindspore.cn/

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3882–3890. ijcai.org.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining answers with entailment trees. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7358–7370, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

James Garson. 2021. Modal Logic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Summer 2021 edition. Metaphysics Research Lab, Stanford University.

Ruixin Hong, Hongming Zhang, Xintong Yu, and Changshui Zhang. 2022. METGEN: A module-based entailment tree generation framework for answer explanation. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1887–1905. Association for Computational Linguistics.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. DAGN: Discourse-aware graph network for logical reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5848–5855, Online. Association for Computational Linguistics.

George Edward Hughes, Max J Cresswell, and Mary Meyerhoff Cresswell. 1996. *A New Introduction to Modal Logic*. Psychology Press.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, Online. Association for Computational Linguistics.

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

Jiaqi Li, Ming Liu, Bing Qin, and Ting Liu. 2022. A survey of discourse parsing. *Frontiers of Computer Science*, 16(5):1–12.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. LogiQA: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3622–3628. ijcai.org.

Charles Lovering, Rohan Jha, Tal Linzen, and Ellie Pavlick. 2021. Predicting inductive biases of pre-trained models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Siru Ouyang, Zhuosheng Zhang, and Hai Zhao. 2021. Fact-Driven logical reasoning. *CoRR*, abs/2105.10334.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An explanation graph generation task for structured commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7716–7740, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1501–1510. ACL.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 46–56. ACL.

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

Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. 2021. ProofWriter: Generating implications, proofs, and abductive statements over natural language. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3621–3634, Online. Association for Computational Linguistics.

Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.

Siyuan Wang, Wanjun Zhong, Duyu Tang, Zhongyu Wei, Zhihao Fan, Daxin Jiang, Ming Zhou, and Nan Duan. 2022. Logic-driven context extension and data augmentation for logical reasoning of text. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1619–1629, Dublin, Ireland. Association for Computational Linguistics.

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

Fangzhi Xu, Jun Liu, Qika Lin, Yudai Pan, and Lingling Zhang. 2022. Logiformer: A two-branch graph transformer network for interpretable logical reasoning. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1055–1065. ACM.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

A Example Source Data from ReClor

Tables 6, 7, 8, and 9 demonstrate the example source data of different reasoning types from the ReClor dataset.

B Guideline for Logical Relation Annotation

Tables 20, 21, and 22 show mappings from natural language patterns to binary logical operators referring to the PDTB3 senses (Webber et al., 2019). For the logical implication, the order of arguments is provided. The three tables are provided to the annotators as their references during annotation. The final annotation is subject to human understanding.

C Indicators of Unary Operators

Table 13 demonstrates the indicators for extracting the unary operators.

D Annotation Interface

The annotation interface is demonstrated in Figure 6.

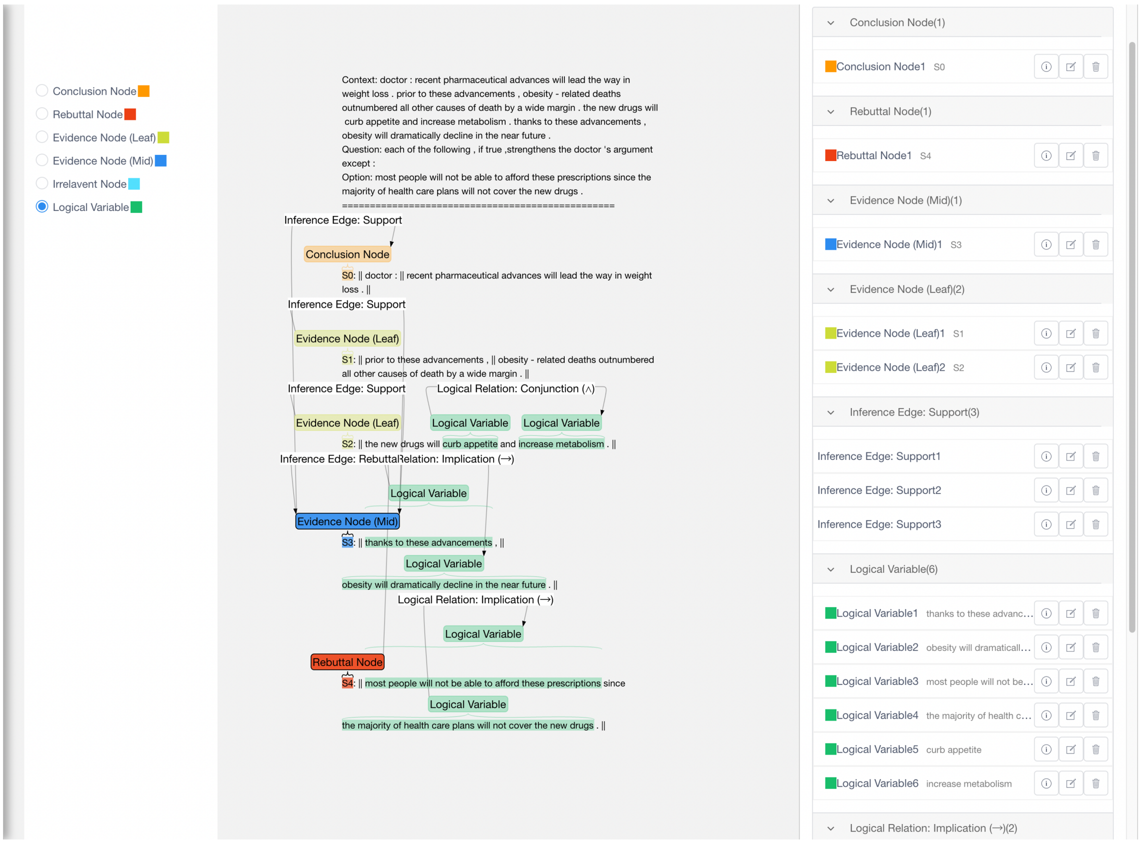
E Model Details

E.1 All-at-Once T5

The input of the All-at-Once T5 is the overall logical passages with sentences and variable denotations. The output is the linearized metagraph as shown in Table 10. The linearized metagraph consists of three parts: the meta structure, node

Reasoning Type: Necessary Assumptions
Context: A recent study showed that people who address problems quickly and directly are significantly less likely to have gum disease than are people who react to problems by refusing to think about them. Since stress can have a negative effect on the immune system, the study’ s results clearly indicate that some forms of gum disease are caused or aggravated by suppression of the immune system. **Question:** The argument requires the assumption that **Options:** A: people who tend to address problems quickly and directly will invariably seek dental care at the first sign of problems B: painful conditions will interfere with a person’s ability to address problems quickly and directly C: people who have highly stressful lives tend to address problems quickly and directly **D:** refusing to think about something troubling contributes to a person’s level of stress

Table 6: Question from ReClor with logical reasoning type: Necessary Assumption. Correct answer option in bold.



Context: doctor : recent pharmaceutical advances will lead the way in weight loss . prior to these advancements , obesity - related deaths outnumbered all other causes of death by a wide margin . the new drugs will curb appetite and increase metabolism . thanks to these advancements , obesity will dramatically decline in the near future .
Question: each of the following , if true ,strengthens the doctor 's argument except :
Option: most people will not be able to afford these prescriptions since the majority of health care plans will not cover the new drugs .
===

Inference Edge: Support

Conclusion Node

S0: || doctor : || recent pharmaceutical advances will lead the way in weight loss . ||

Inference Edge: Support

Evidence Node (Leaf)

S1: || prior to these advancements , || obesity - related deaths outnumbered all other causes of death by a wide margin . ||

Inference Edge: Support

Logical Relation: Conjunction (∧)

Evidence Node (Leaf)

Logical Variable

Logical Variable

S2: || the new drugs will curb appetite and increase metabolism . ||

Inference Edge: RebuttaRelation: Implication (→)

Logical Variable

Evidence Node (Mid)

S3: || thanks to these advancements , ||

Logical Variable

obesity will dramatically decline in the near future . ||

Logical Relation: Implication (→)

Logical Variable

Rebuttal Node

S4: || most people will not be able to afford these prescriptions since

Logical Variable

the majority of health care plans will not cover the new drugs . ||

Conclusion Node(1)
🟧Conclusion Node1 S0
Rebuttal Node(1)
🟥Rebuttal Node1 S4
Evidence Node (Mid)(1)
🟦Evidence Node (Mid)1 S3
Evidence Node (Leaf)(2)
🟨Evidence Node (Leaf)1 S1
🟨Evidence Node (Leaf)2 S2
Inference Edge: Support(3)
Inference Edge: Support1
Inference Edge: Support2
Inference Edge: Support3
Logical Variable(6)
🟩Logical Variable1 thanks to these advanc…
🟩Logical Variable2 obesity will dramaticall…
🟩Logical Variable3 most people will not be…
🟩Logical Variable4 the majority of health c…
🟩Logical Variable5 curb appetite
🟩Logical Variable6 increase metabolism
Logical Relation: Implication (→)(2)

Figure 6: The annotation interface.

Reasoning Type: Sufficient Assumptions
Context: In Europe, schoolchildren devote time during each school day to calisthenics. North American schools rarely offer a daily calisthenics program. Tests prove that North American children are weaker, slower, and shorter-winded than European children. We must conclude that North American children can be made physically fit only if they participate in school calisthenics on a daily basis. **Question:** Which one of the following is assumed in the passage? **Options:** **A:** School calisthenics are an indispensable factor in European children's superior physical fitness. B: All children can be made physically fit by daily calisthenics. C: Superior physical fitness produces superior health. D: North American children can learn to eat a more nutritious diet as well as to exercise daily.

Table 7: Question from ReClor with logical reasoning type: Sufficient Assumption. Correct answer option in bold.

Reasoning Type: Strengthen
Context: Skeletal remains of early humans indicate clearly that our ancestors had fewer dental problems than we have. So, most likely, the diet of early humans was very different from ours. **Question:** Which one of the following, if true, most strengthens the argument? **Options:** A: Skeletal remains indicate that some early humans had a significant number of cavities. B: A healthy diet leads to healthy teeth. **C:** Diet is by far the most significant factor contributing to dental health. D: Early humans had a shorter average life span than we do, and the most serious dental problems now tend to develop late in life.

Table 8: Question from ReClor with logical reasoning type: Strengthen. Correct answer option in bold.

formula, and sentence certainty (denoted with $graph$, $formula$, and $degree$, respectively). For the meta structure, we use the semicolon to connect different edges. For the node formula, we map the operator to word using the mapping { □: [necessary], ◇: [possible], ¬: [negative], ∧: [and], ∨: [or], →: [entail] }. We use the semicolon to connect the triples for the same sentence. We connect the certainties and formulae of sentences with "|".

Reasoning Type: Weaken
Context: Many people suffer an allergic reaction to sulfites, including those that are commonly added to wine as preservatives. However, since there are several winemakers producing wine to which no sulfites are added, those who would like to drink wine but are allergic to sulfites can drink these wines without risking an allergic reaction to sulfites. **Question:** Which of the following, if true, most seriously weakens the argument? **Options:** **A:** Sulfites occur naturally in most wine. B: The sulfites that can produce an allergic reaction are also commonly found in beverages other than wine. C: Wine without added sulfites sometimes becomes undrinkable even before the wine is sold to consumers. D: Apart from sulfites, there are other substances commonly present in wine that can trigger allergic reactions.

Table 9: Question from ReClor with logical reasoning type: Weaken. Correct answer option in bold.

Input: sent1: v1: measurements of the motion of the planet uranus seem to show uranus being tugged by a force pulling it away from the sun and the inner planets . sent2: v1: neptune and pluto , v2: the two known planets whose orbits are farther from the sun than is the orbit of uranus , v3: do not have enough mass to exert the force that the measurements indicate . sent3: v1: therefore , v2: in addition to the known planets , v3: there must be at least one planet in our solar system that we have yet to discover . sent4: v1: there is a belt of comets beyond the orbit of pluto with powerful gravitational pull .
Output: \$graph\$ sent1 -> sent3; sent2 -> sent3; sent4 => sent2; \$formula\$ sent3: v2 [and] [necessary] v3; \$degree\$ sent1: contingent | sent2: contingent | sent3: necessary | sent4: contingent

Table 10: Example of the metagraph generation task. "->" denotes the *support* relation, and "=>" denotes the *rebut* relation. The readable triple in `sent3` formulas is $v_2 \wedge \Box v_3$.

We train the All-at-Once T5 model with a batch size of 32 and a learning rate of 1e-5 for 300 epochs.

E.2 Multitask T5

The Multitask T5 decomposes the whole generation task into three sub-tasks: meta structure generation, formula generation, and certainty prediction. We train a single T5 model on these three sub-tasks simultaneously. We follow Raffel et al. (2020) to add a task-specific prefix to the input before feeding it to the model. Table 11 shows some specific input and output examples of each sub-task. We train the Multitask T5 model with a batch size of 32 and a learning rate of 1e-5 for 300 epochs. We use the examples-proportional mixing (Raffel et al., 2020) and simply concatenate the data for all sub-tasks as the training data for Multitask T5.

Task: Meta structure generation **Input:** `GRAPH:` sent1: to reduce waste of raw materials , the government of sperland is considering requiring household appliances to be broken down for salvage when discarded . [AND] imposing the fee at the time of salvage would reduce waste more effectively , however , because consumers tend to keep old appliances longer if they are faced with a fee for discarding them . sent2: to cover the cost of salvage , the government is planning to charge a fee , which would be imposed when the appliance is first sold . sent4: increasing the cost of disposing of an appliance properly increases the incentive to dispose of it improperly . **Output:** sent4 => sent1; sent1 -> sent2;
Task: Formula generation **Input:** `FORMULAE:` v1: grammarians have for years condemned as ungrammatical the english phrase " between you and i " , insisting that the correct phrasing is " between you and me , " with v2: the objective case after v3: a preposition . v4: such condemnations , however , are obviously unfounded , because v5: shakespeare himself , in the merchant of venice , wrote , " all debts are cleared between you and i. " **Output:** [necessary] v5 [entail] [necessary] v4; v2 [entail] v3;
Task: Certainty prediction **Input:** `DEGREE:` it was formerly believed that prehistoric homo sapiens ancestors of contemporary humans interbred with neanderthals , but dna testing of a neanderthal ' s remains indicates that this is not the case . **Output:** impossible

Table 11: Input and output examples of all the sub-tasks of Multitask T5.

E.3 MetGen

MetGen (Hong et al., 2022) is a module-based framework for structured explanation generation.

Modules. We use two types of modules: the conclusion module and the rebuttal module. The conclusion module takes two sentences as input (e.g., `sent1:... sent2:...`) and outputs the inference relation type between them (e.g., `sent1 -> sent2` or `sent2 -> sent1`). If there is no conclusive relationship between the two input sentences, the module would output the word `none`. The rebuttal module is defined similarly.

Controller. The controller decides the reasoning direction based on the current reasoning state. Specifically, given the current partially metagraph and all the sentences, the controller predicts which combinations of two sentences (e.g., `sent1 sent2`) should be considered in the next step. If the current proof is complete, the controller would output the word `done`.

Reasoning Process. MetGen generates the metagraphs in an iterative manner. It iteratively repeats the following reasoning iteration to grow the graph until either the controller returns `done` or the maximum number of iteration steps is reached. It takes several iterations before completing the generation. In each iteration, MetGen generates one step. It first uses the controller to predict some possible sentence combinations. Then, each combination is sent to the reasoning modules to generate candidate steps that indicate the detailed inference relation type between them. The candidate step with the highest score (the lowest perplexity) is picked for the next iteration.

Implementations. To compare with other methods under the same number of parameters, we implement MetGen using a single T5 model. Table 12 shows some input and output examples of MetGen. The MetGen is trained on five sub-tasks simultaneously: controller task, conclusion module task, rebuttal module task, formula generation task, and certainty prediction tasks. We train the MetGen model with a batch size of 32 and a learning rate of 1e-5 for 300 epochs. We set the maximum number of iteration steps as 3.

Task: Controller **Input:** `CONTROL`: proof: sent1 -> sent3; context: sent1: measurements of the motion of the planet uranus seem to show uranus being tugged by a force pulling it away from the sun and the inner planets . sent2: neptune and pluto , the two known planets whose orbits are farther from the sun than is the orbit of uranus , do not have enough mass to exert the force that the measurements indicate . sent3: therefore , in addition to the known planets , there must be at least one planet in our solar system that we have yet to discover . sent4: there is a belt of comets beyond the orbit of pluto with powerful gravitational pull . **Output:** sent2 sent3
Task: Conclusion Module **Input:** `CONCLUSION`: sent2: the dna of contemporary humans is significantly different from that of the neanderthal. sent3: the dna of prehistoric homo sapiens ancestors of contemporary humans was not significantly more similar to that of neanderthals than is the dna of contemporary humans. **Output:** sent3 -> sent2
Task: Rebuttal Module **Input:** `REBUTTAL`: sent1: recent unexpectedly heavy rainfalls in the metropolitan area have filled the reservoirs and streams ; water rationing , therefore , will not be necessary this summer . sent2: the water company 's capacity to pump water to customers has not kept up with the increased demand created by population growth in the metropolitan area . **Output:** sent2 => sent1

Table 12: Input and output examples of the controller and modules of MetGen. MetGen is also trained with the formula generation and certainty prediction tasks.

E.4 Experimental Details

We use the pre-trained models from `HuggingFace Transformers`[7]. We use the Adafactor optimizer (Shazeer and Stern, 2018). We run the experiments based on T5-large 3 times with different random seeds and report the average performances. The experiments based on T5-11b are run only once considering the computational cost.

F Evaluation Metrics

Meta structure: Does the predicted metagraph use the correct sentences and have the correct structure? For meta nodes, we report a node F1 score by comparing the set of sentences used in the predicted and gold metagraph. For meta structure, we decompose the metagraph into one-premise steps (e.g., `sent1 -> sent2`). We compare the set of steps in the predicted and gold metagraph and report the step F1 score. A predicted step is correct if its premise, conclusion, and step type match the gold one. The AllCorrect score is 1 if the F1 is 1, 0 otherwise.

Formula: Does the predicted metagraph have the correct internal structure of meta nodes? For each sentence, we measure the formula F1 score by comparing all formulae in the predictions and gold annotations. A predicted formula is considered correct if its certainty operators, binary operators, and variables match the gold one. For the certainty operators, we reduce them to the standard form (one of the five degree-of-certainty forms listed in Table 1) before comparison. For the binary operator, we consider its symmetry. For example, $\neg \mathrm{v}_1 \land \Diamond \mathrm{v}_2$ is equivalent to $\Diamond \mathrm{v}_2 \land \neg \mathrm{v}_1$, but $\neg \mathrm{v}_1 \rightarrow \Diamond \mathrm{v}_2$ is not equivalent to $\Diamond \mathrm{v}_2 \rightarrow \neg \mathrm{v}_1$. The AllCorrect score is 1 if the formula F1 is 1, 0 otherwise. Since each sample contains multiple sentences, we average the formula F1 scores of all sentences in the sample as the formula F1 score for this sample.

Certainty: Are the certainties of the sentence correct? For each sample, we compute the accuracy of the predicted certainties of the sentences. The AllCorrect score is 1 if the accuracy is 1, and 0 otherwise. We report the accuracy and AllCorrect score of the testing dataset, which is the average accuracy and AllCorrect score of all samples in

[7] https://github.com/huggingface/transformers

Negation (¬)	"no", "not", "none", "nobody", "nothing", "neither", "nor", "nowhere", "never", "hardly", "scarcely", "barely", "doesn’t", "isn’t", "wasn’t", "shouldn’t", "wouldn’t", "couldn’t", "won’t", "can’t", "don’t", "impossible"
Box (□)	"necessarily", "must", "definitely", "certainly", "clearly", "obviously", "undoubtedly", "surely", "will", "all", "every", "always"
Diamond (◇)	"likely", "approximately", "possibly", "perhaps", "probably", "maybe", "few", "may", "might", "could", "many", "most", "some", "numerous", "countless", "majority", "often", "frequently", "commonly", "usually", "sometimes", "repeatedly", "appears", "seems", "suggests", "indicates"

Table 13: Indicators of unary operators.

the dataset. Due to the unbalance of the certainty labels, we gather the predictions for all sentences in the dataset (ignoring which sample the sentence comes from) and report the macro-F1 score.

Overall: The overall AllCorrect score of a predicted metagraph is 1 only if all of the meta structure, formulae, and certainties are correct. This is a strict metric since any error would result in a score of 0.

G Detailed Analysis Results

Table 14 present the detailed performance of different inference steps and different operators. Table 15 and Table 16 shows the detailed results with different ratios of training data.

H Error Cases

Examples of each error type are shown in Tables 17, 18, and 19.

	Support		Rebut		Operators						
	F1	All	F1	All	→	∧	∨	¬	□	◇	N/A
Once (large)	48.9	15.8	53.3	53.0	25.4	23.8	2.1	3.2	3.4	7.5	83.1
Multitask (large)	56.4	22.5	57.0	56.8	52.2	38.2	27.1	13.3	6.8	13.7	98.4
MetGen (large)	60.9	26.8	59.8	59.3	58.8	41.9	33.3	14.3	11.4	24.9	96.3
Once (11b)	57.1	27.0	59.5	59.5	42.6	34.5	18.8	3.2	6.3	13.5	93.6
Multitask (11b)	62.8	34.0	64.2	63.5	52.0	41.9	31.2	7.1	9.4	15.3	99.3
MetGen (11b)	65.6	32.5	69.7	69.0	60.8	52.5	43.8	12.7	15.6	18.8	97.2

Table 14: Performance on two types of inference step. Per-operator F1 scores.

		Component 1				Component 2		Component 3			Overall
		Node		Step		Formulae		Certainty			
	Ratio	F1	All	F1	All	F1	All	Acc	All	F1*	All
Once (large)	50%	89.2	50.5	39.0	14.0	52.5	51.6	73.2	34.0	22.4	3.6
	20%	86.6	45.5	36.1	11.0	44.6	43.7	70.7	29.0	18.3	3.5
	10%	88.8	49.0	30.1	6.5	42.0	41.8	65.6	25.5	18.2	1.2
	5%	82.3	26.0	20.0	0.5	40.2	40.0	66.8	20.0	14.7	0.5
	2%	71.0	5.5	14.6	0.0	31.8	31.4	65.3	18.0	13.7	0.0
	1%	70.7	2.5	13.6	0.5	38.9	38.8	64.7	14.5	15.8	0.0
Multitask (large)	50%	94.1	68.5	55.2	22.0	78.9	76.9	84.9	57.5	50.8	10.5
	20%	92.4	64.0	44.7	16.0	70.3	68.7	81.0	47.0	22.9	5.0
	10%	56.1	3.5	13.3	1.5	71.1	70.2	80.3	47.0	23.0	0.3
	5%	38.9	2.0	5.7	0.5	65.9	65.0	80.2	45.0	18.7	0.0
	2%	31.1	2.5	3.9	1.0	61.5	60.6	80.7	46.0	18.6	0.3
	1%	8.6	2.0	2.0	1.0	56.7	56.4	80.8	46.0	18.0	0.0
MetGen (large)	50%	94.3	67.0	53.1	17.5	75.5	74.1	83.5	56.5	31.1	9.2
	20%	92.4	64.0	44.7	16.0	70.3	68.7	81.0	47.0	22.9	5.0
	10%	56.1	3.5	13.3	1.5	71.1	70.2	80.3	47.0	23.0	0.3
	5%	38.9	2.0	5.7	0.5	65.9	65.0	80.2	45.0	18.7	0.0
	2%	31.1	2.5	3.9	1.0	61.5	60.6	80.7	46.0	18.6	0.3
	1%	8.6	2.0	2.0	1.0	56.7	56.4	80.8	46.0	18.0	0.0

Table 15: Detailed results on different ratios of MetaLogic training data. *: macro-F1.

		Support		Rebut		Operators						
	Ratio	F1	All	F1	All	→	∧	∨	¬	□	◇	N/A
Once (large)	50%	42.2	16.0	47.8	47.5	19.1	13.8	0.0	0.0	0.0	2.8	79.4
	20%	38.4	12.0	49.3	49.0	11.3	6.7	0.0	3.2	2.1	6.9	71.6
	10%	31.7	7.5	50.3	50.0	5.9	5.6	0.0	9.5	2.1	0.0	69.7
	5%	20.4	1.5	26.3	26.0	5.5	1.8	0.0	0.0	0.0	0.0	68.8
	2%	17.9	1.0	15.5	15.5	3.7	5.7	0.0	0.0	0.0	0.0	53.4
	1%	17.2	1.5	12.5	12.5	4.3	1.4	0.0	0.0	2.1	2.1	65.7
Multitask (large)	50%	59.2	24.5	60.3	60.0	57.2	47.4	37.5	15.1	17.2	26.0	97.2
	20%	48.9	18.5	52.5	52.5	36.2	22.3	18.8	0.0	0.0	3.8	98.3
	10%	16.9	6.5	56.0	56.0	36.1	22.3	6.2	4.8	0.0	2.1	99.5
	5%	7.6	3.0	58.0	58.0	30.2	14.6	12.5	0.0	1.6	8.3	96.2
	2%	4.8	2.5	58.0	58.0	34.5	12.7	0.0	0.0	3.6	4.9	87.5
	1%	4.2	3.5	59.0	59.0	25.6	5.4	0.0	0.0	2.1	2.8	85.3
MetGen (large)	50%	58.1	20.5	58.5	58.0	47.8	34.5	25.0	12.7	1.6	11.1	98.6
	20%	53.3	20.5	56.2	55.5	43.1	28.5	18.8	0.0	3.1	7.9	97.6
	10%	55.3	20.5	55.8	55.5	35.5	26.2	6.2	4.8	0.0	4.2	99.3
	5%	50.5	16.0	55.8	55.5	32.1	23.0	6.2	0.0	0.0	3.5	93.9
	2%	40.2	3.5	58.5	58.5	28.3	13.6	0.0	0.0	1.6	6.2	91.0
	1%	35.0	3.0	59.0	59.0	23.4	3.0	0.0	0.0	2.1	2.8	89.1

Table 16: Inference step and operator performances on different ratios of MetaLogic training data.

G1: Incorrect Inference Type

Passage: sent1: for similar cars and drivers , automobile insurance for collision damage has always cost more in greatport than in fairmont . [AND] police studies , however , show that cars owned by greatport residents are , on average , slightly less likely to be involved in a collision than cars in fairmont . sent3: clearly , therefore , insurance companies are making a greater profit on collision - damage insurance in greatport than in fairmont . sent4: repairing typical collision damage does not cost more in greatport than in fairmont .
Gold: sent4 -> sent3; sent1 -> sent3;
Pred: sent1 -> sent3; sent4 => sent3;

G2: Incorrect Rebuttal

Passage: sent1: there should be a greater use of gasohol . sent2: gasohol is a mixture of alcohol and gasoline , and has a higher octane rating and fewer carbon monoxide emissions than straight gasoline . [AND] burning gasohol adds no more carbon dioxide to the atmosphere than plants remove by photosynthesis . sent4: cars burn on the average slightly more gasohol per kilometer than they do gasoline .
Gold: sent2 -> sent1; sent4 => sent2;
Pred: sent4 -> sent1; sent2 -> sent1;

G3: Incorrect Conclusion

Passage: sent1: healthy lungs produce a natural antibiotic that protects them from infection by routinely killing harmful bacteria on airway surfaces . [AND] people with cystic fibroses , however , are unable to fight off such bacteria , even though their lungs produce normal amounts of the antibiotic . sent3: since the fluid on airway surfaces in the lungs of people with cystic fibrosis has an abnormally high salt concentration , scientists hypothesize that in high salt environments the antibiotic becomes ineffective at killing harmful bacteria . sent4: the lungs of people who suffer from cystic fibrosis are unable to fight off harmful bacteria even when the salt concentration is reduced to levels typical of healthy lungs .
Gold: sent3 -> sent1; sent4 => sent3;
Pred: sent4 -> sent3; sent1 -> sent3;

G4: Incorrect Inference Step

Passage: sent1: spokesperson : the major school lunch vendors recently agreed to stop selling high - calorie beverages in elementary and middle schools because studies show that children of ages 7 to 8 who substitute one low - calorie beverage for one high - calorie soft drink in their daily diets will , on average , weigh 20 pounds less than they would have by the time they reach high school . sent2: since only low - calorie beverages will be sold in schools , within six to eight years , we can expect to see a reduction in the percentage of overweight high - school children . sent3: elementary and middle school students who used to buy high - calorie soft drinks at school will not bring them to school or drink extra high - calorie beverages at home as a substitute .
Gold: sent3 -> sent2; sent1 -> sent2;
Pred: sent3 -> sent2;

G5: Other Structural Mismatch

Passage: sent1: employer : in the current economic climate , the best way to run a business is to pay employees the least amount possible to do the job . sent2: the supply of labor is far outpacing demand since the number of college graduates increases every year and the average age of retirement is also increasing . [AND] applicants will typically take the first job offer on the table , and any employee who demands a raise can be easily replaced from the labor pool . sent4: even if the employee is unhappy , he or she will often remain on the job due to the competition in the job market . [AND] keeping payroll costs low allows more resources to be devoted to innovation , delivering a higher quality product to customers . sent6: automation is the leading cause for unemployment .
Gold: sent4 -> sent1; sent2 -> sent4;
Pred: sent6 => sent1; sent2 -> sent1;

Table 17: Error cases for meta graph structure.

F1: Incorrect Logical Variable

Sentence: v1: legislators considering a proposed law for which they have v2: repugnance or v3: enthusiasm v4: do not consider the consequences that it will actually have .
Gold: v3 [or] v2;
Pred: v2 [or] v3; v4 [entail] v1;

F2: Incorrect Unary Operator

Sentence: v1: auditor : xyz , v2: a construction company , purchased 20 new trucks 3 years ago , and v3: there is no record of any of those trucks being sold last year .
Gold: v2 [and] v3;
Pred: v1 [and] v3;

F3: Incorrect Binary Operator

Sentence: v1: travaillier corporation has recently hired employees with experience in the bus tour industry v2: its executives have also been negotiating with charter bus companies that subcontract with bus tour companies . [AND] but v3: travaillier has traditionally focused on serving consumers who travel primarily by air , and v4: marketing surveys show that travaillier ’ s traditional consumers have not changed their vacation preferences .
Gold: v1 [and] v2; v3 [and] v4;
Pred: v3 [and] v4; v2 [entail] v1;

F4: Incorrect Implication Direction

Sentence: v1: now some politicians are saying that , in order to v2: cause another similarly sized increase in exports , v3: the government should allow the pundra to become weak again .
Gold: v3 [entail] v2;
Pred: v2 [entail] v3;

Table 18: Error cases for the formulae.

C1: Incorrect Polarity
Sentence: the chemistry department ’s funding for basic science research is not likely to increase if its funding from sources other than profit - driven institutions does not increase . **Gold:** impossible **Pred:** possible
C2: Other Polarities to Contingent
Sentence: if legislators are to enact laws that benefit constituents , they must be sure to consider what the consequences of enacting a proposed law will actually be . [AND] concerned primarily with advancing their own political careers , legislators present legislation in polemical terms ; this arouses in their colleagues either repugnance or enthusiasm for the legislation . **Gold:** necessary **Pred:** contingent
C3: Contingent to Other Polarities
Sentence: making decisions about patterns of work organization , resource allocation , and location of industry is not the core of a public official ’s job . **Gold:** contingent **Pred:** unnecessary
C4: Unresolved Certainty
Sentence: the link between jogging and certain structural disorders appears to be a causal one . **Gold:** contingent **Pred:** causal

Table 19: Error cases for certainty.

Connectives	**Binary Operator**	**Senses**
about	$A \rightarrow B$	Contingency.Cause.Reason
A accordingly B. A; accordingly B. A. B accordingly.	$A \rightarrow B$	Contingency.Cause.Result
after	$A \rightarrow B$	Temporal.Asynchronous.Succession Contingency.Cause.Reason
afterward	$A \rightarrow B$	Temporal.Asynchronous.Precedence
afterwards	$A \rightarrow B$	Temporal.Asynchronous.Precedence
B, as A. As A, B.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg1-as-detail Expansion.Manner.Arg2-asmanner Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Asynchronous.Succession
as a result	$A \rightarrow B$	Contingency.Cause.Result
B, as if A. As if A, B.	$A \rightarrow B$	Comparison.Concession.Arg2-as-denier Comparison.Similarity Expansion.Manner.Arg2-asmanner Expansion.Instantiation.Arg1-as-instance Expansion.Manner.Arg2-asmanner
Because of A, B. B because of A.	$A \rightarrow B$	Contingency.Cause.Reason
Because A, B. B because A.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason Contingency.Condition+SpeechAct
B, as long as A. As long as A, B.	$A \rightarrow B$	Contingency.Condition.Arg2 -as-cond
Before B, A. A before B.	$A \rightarrow B$	Temporal.Asynchronous.Precedence Temporal.Asynchronous.Succession
By A, B. B by A.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Expansion.Manner.Arg2-asmanner Contingency.Cause.Reason Expansion.Manner.Arg2-asmanner Contingency.Cause.Reason Contingency.Condition.Arg2-as-cond Expansion.Manner.Arg2-asmanner Contingency.Condition.Arg2-as-cond Contingency.Purpose.Arg1as-goal Expansion.Manner.Arg2-asmanner Expansion.Level-of-detail.Arg2-as-detail Expansion.Manner.Arg2-asmanner
by then	$A \rightarrow B$	Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Asynchronous.Succession
A consequently B.	$A \rightarrow B$	Contingency.Cause.Result
B depending on A. Depending on A, B.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
B depending upon A. Depending upon A, B.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
B, due to A. Due to A, B.	$A \rightarrow B$	Contingency.Cause.Reason
B. Earlier, A. B, A earlier.	$A \rightarrow B$	Temporal.Asynchronous.Su ccession
B even after A. Even after A, B.	$A \rightarrow B$	Temporal.Asynchronous.Succession Comparison.Concession.Arg1-as-denier
A, finally, B.	$A \rightarrow B$	Temporal.Asynchronous.Precedence Contingency.Cause.Result Temporal.Asynchronous.Precedence
B, for A. For A, B.	$A \rightarrow B$	Comparison.Concession.Arg1-as-denier Contingency.Cause.Reason Contingency.Cause.Result Contingency.Condition.Arg2 -as-cond Contingency.Purpose.Arg2as-goal Expansion.Level-of-detail.Arg2-as-detail
B, for example A.	$A \rightarrow B$	Expansion.Instantiation.Arg2-as-instance

B, for instance A.	$A \rightarrow B$	Expansion.Instantiation.Arg 2-as-instance
B, from A.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason Contingency.Condition.Arg2-as-cond Contingency.Cause.Reason Expansion.Manner.Arg2-asmanner Contingency.Cause.Reason Expansion.Substitution.Arg1-as-subst
Given A, B. B, given A.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason
A, hence B.	$A \rightarrow B$	Contingency.Cause.Result
If A, B. B, if A.	$A \rightarrow B$	Comparison.Concession+SpeechAct.Arg2-as-denier+SpeechAct Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier Comparison.Contrast Contingency.Condition+SpeechAct Contingency.Condition.Arg2-as-cond Expansion.Level-of-detail.Arg2-as-detail Contingency.Condition.Arg2-as-cond Temporal.Synchronous Contingency.Condition.Arg2-as-cond
B if and when A.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond Temporal.Synchronous Contingency.Condition.Arg2-as-cond
B if and when A.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond Temporal.Synchronous Contingency.Condition.Arg2-as-cond
B if only A.	$A \rightarrow B$	Comparison.Concession.Arg2-as-denier Contingency.Condition.Arg2-as-cond Contingency.Purpose.Arg2as-goal
If A then B.	$A \rightarrow B$	Contingency.Condition+SpeechAct Contingency.Condition.Arg2-as-cond
B in A.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason Expansion.Manner.Arg2-asmanner Contingency.Cause.Reason Contingency.Condition.Arg2-as-cond Expansion.Manner.Arg2-asmanner Contingency.Condition.Arg2-as-cond Contingency.Purpose.Arg2as-goal Expansion.Instantiation.Arg1-as-instance Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail Expansion.Manner.Arg2-asmanner Temporal.Synchronous Contingency.Purpose.Arg2as-goal Temporal.Synchronous Expansion.Level-of-detail.Arg1-as-detail Temporal.Synchronous
B in case A.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
B. In fact, A. B, A in fact.	$A \rightarrow B$	Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail
B in order A.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond Contingency.Purpose.Arg2as-goal
B, in particular A.	$A \rightarrow B$	Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg2-as-detail
A, in short, B.	$A \rightarrow B$	Expansion.Level-of-detail.Arg1-as-detail
A, in sum, B.	$A \rightarrow B$	Expansion.Level-of-detail.Arg1-as-detail
B, in that A.	$A \rightarrow B$	Expansion.Level-of-detail.Arg2-as-detail
A, in the end B.	$A \rightarrow B$	Contingency.Cause.Result Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail Temporal.Asynchronous.Pre cedence

B, indeed A. B, A indeed.	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason Contingency.Cause.Result Expansion.Conjunction Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg1-as-detail Expansion.Level-of-detail.Arg2-as-detail
B insofar as A. Insofar as A, B.	$A \rightarrow B$	Contingency.Cause.Reason Expansion.Level-of-detail.Arg2-as-detail
A, B later.	$A \rightarrow B$	Temporal.Asynchronous.Precedence
A, B later on.	$A \rightarrow B$	Temporal.Asynchronous.Precedence
B, more accurately, A.	$A \rightarrow B$	Expansion.Substitution.Arg2-as-subst
A, next B.	$A \rightarrow B$	Temporal.Asynchronous.Precedence Expansion.Conjunction Temporal.Asynchronous.Precedence
B, not only because of A.	$A \rightarrow B$	Contingency.Cause.Reason
Now that A, B.	$A \rightarrow B$	Contingency.Cause.Reason Temporal.Asynchronous.Precedence Contingency.Cause.Reason Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Synchronous Contingency.Cause.Reason Temporal.Synchronous
B on A.	$A \rightarrow B$	Contingency.Cause.Reason
Once A, B. B, once A.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Asynchronous.Succession Contingency.Condition.Arg2as-cond Temporal.Asynchronous.Succession
B only if A.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
B previously A.	$A \rightarrow B$	Temporal.Asynchronous.Succession Comparison.Contrast Temporal.Asynchronous.Succession
B, since A. Since A, B.	$A \rightarrow B$	Contingency.Cause.Reason Temporal.Asynchronous.Precedence Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Asynchronous.Succession
B since before A.	$A \rightarrow B$	Temporal.Asynchronous.Succession
A, so B.	$A \rightarrow B$	Contingency.Cause+Belief.Result+Belief Contingency.Cause.Result Contingency.Purpose.Arg2as-goal
So as A, B.	$A \rightarrow B$	Contingency.Purpose.Arg2as-goal
B so long as A. So long as A, B.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
A so that B.	$A \rightarrow B$	Contingency.Cause.Result Contingency.Purpose.Arg2as-goal
B, specifically, A.	$A \rightarrow B$	Expansion.Level-of-detail.Arg2-as-detail
A subsequently B.	$A \rightarrow B$	Temporal.Asynchronous.Precedence
B such as A.	$A \rightarrow B$	Expansion.Instantiation.Arg2-as-instance
B, that is A.	$A \rightarrow B$	Expansion.Equivalence Expansion.Level-of-detail.Arg2-as-detail
A then B.	$A \rightarrow B$	Contingency.Cause.Result Expansion.Conjunction Contingency.Cause.Result Contingency.Condition.Arg1-as-cond Temporal.Asynchronous.Precedence Contingency.Cause.Result Temporal.Asynchronous.Precedence
A thereafter B.	$A \rightarrow B$	Temporal.Asynchronous.Precedence
A thereby B.	$A \rightarrow B$	Contingency.Cause.Result Expansion.Manner.Arg1-asmanner
A therefore B.	$A \rightarrow B$	Contingency.Cause.Result
A thus B.	$A \rightarrow B$	Contingency.Cause+Belief.Result+Belief Contingency.Cause.Result
B till A.	$A \rightarrow B$	Contingency.Negative-condition.Arg2-as-negCond Temporal.Asynchronous.Precedence

A ultimately B.	$A \rightarrow B$	Contingency.Cause.Reason Expansion.Conjunction Temporal.Asynchronous.Precedence Contingency.Cause.Result Temporal.Asynchronous.Precedence
B untill A. Untill A, B.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond Temporal.Asynchronous.Precedence Temporal.Asynchronous.Succession Contingency.Condition.Arg2-as-cond Temporal.Asynchronous.Su ccession
B upon A. Upon A, B.	$A \rightarrow B$	Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Synchronous Contingency.Cause.Reason Temporal.Synchronous
B, when A. When A, B.	$A \rightarrow B$	Contingency.Cause.Reason Contingency.Condition+SpeechAct Contingency.Condition.Arg2-as-cond Expansion.Level-of-detail.Arg2-as-detail Contingency.Condition.Arg2-as-cond Expansion.Manner.Arg2-asmanner Temporal.Asynchronous.Precedence Contingency.Condition.Arg2-as-cond Temporal.Asynchronous.Precedence Temporal.Asynchronous.Succession Contingency.Cause+Belief.Reason+Belief Temporal.Asynchronous.Succession Contingency.Cause.Reason Temporal.Asynchronous.Succession Contingency.Cause.Result Temporal.Asynchronous.Succession Contingency.Condition+SpeechAct Temporal.Asynchronous.Succession Contingency.Condition.Arg2-as-cond Temporal.Asynchronous.Su ccession
B when and if A.	$A \rightarrow B$	Temporal.Asynchronous.Succession Contingency.Condition.Arg2-as-cond
B whenever A. Whenever A, B.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
B, where A. Where A, B.	$A \rightarrow B$	Contingency.Condition.Arg2-as-cond
B, with A. With A, B	$A \rightarrow B$	Contingency.Cause+Belief.Reason+Belief Contingency.Cause.Reason Expansion.Level-of-detail.Arg2-as-detail Contingency.Cause.Reason Contingency.Condition.Arg2-as-cond Expansion.Instantiation.Arg2-as-instance Expansion.Level-of-detail.Arg2-as-detail Expansion.Manner.Arg2-asmanner
B, without A. Without A, B.	$A \rightarrow B$	Contingency.Cause.Reason Contingency.Cause.Result Expansion.Level-of-detail.Arg2-as-detail Expansion.Manner.Arg2-asmanner

Table 20: Mapping from connectives to logical implication ($\rightarrow$), according to PDTB senses.

Connectives	Binary Operator	Senses
A, and B.	$A \wedge B$	Comparison.Concession+SpeechAct.Arg2-as-denier+SpeechAct Comparison.Contrast Contingency.Cause+SpeechAct.Result+SpeechAct Contingency.Cause.Reason Contingency.Cause.Result Expansion.Conjunction Contingency.Cause.Result Contingency.Condition.Arg1-as-cond Contingency.Purpose.Arg2as-goal Expansion.Conjunction Expansion.Level-of-detail.Arg2-as-detail Expansion.Manner.Arg2-asmanner
additionally	$A \wedge B$	Expansion.Conjunction
Albeit A, B. B, albeit A.	$A \wedge B$	Comparison.Concession.Arg2-as-denier
along with	$A \wedge B$	Expansion.Conjunction
also	$A \wedge B$	Expansion.Conjunction Temporal.Synchronous
although	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Exception.Arg2-as-excpt Temporal.Synchronous Comparison.Contrast
as long as	$A \wedge B$	Temporal.Synchronous Contingency.Condition.Arg2-as-cond Temporal.Synchronous
as much as	$A \wedge B$	Comparison.Concession.Arg1-as-denier Expansion.Conjunction Expansion.Substitution.Arg2-as-subst
as soon as	$A \wedge B$	Temporal.Asynchronous.Succession Temporal.Synchronous
as though	$A \wedge B$	Comparison.Similarity Expansion.Manner.Arg2-asmanner Comparison.Similarity Expansion.Level-of-detail.Arg2-as-detail
as well	$A \wedge B$	Comparison.Similarity Expansion.Conjunction
as well as	$A \wedge B$	Expansion.Conjunction
as	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Contrast Comparison.Similarity Temporal.Synchronous Comparison.Contrast Temporal.Synchronous Comparison.Similarity Temporal.Synchronous Contingency.Cause+Belief.Reason+Belief Temporal.Synchronous Contingency.Cause.Reason Temporal.Synchronous
at the same time	$A \wedge B$	Temporal.Synchronous
before and after	$A \wedge B$	Temporal.Asynchronous.Precedence Temporal.Asynchronous.Succession
besides	$A \wedge B$	Expansion.Conjunction
A, beyond B. Beyond B, A.	$A \wedge B$	Expansion.Conjunction
both A and B.	$A \wedge B$	Expansion.Conjunction

but	$A \wedge B$	Comparison.Concession+SpeechAct.Arg2-as-denier+SpeechAct Comparison.Concession.Arg2-as-denier Comparison.Contrast Contingency.Cause+SpeechAct.Reason+SpeechAct Contingency.Cause.Reason Comparison.Concession.Arg2-as-denier Expansion.Conjunction Expansion.Exception.Arg2-as-excpt Temporal.Synchronous Comparison.Contrast
A but also B.	$A \wedge B$	Expansion.Conjunction
A but then B.	$A \wedge B$	Comparison.Concession.Arg2-as-denier
A but then again B.	$A \wedge B$	Comparison.Concession.Arg2-as-denier
by comparison	$A \wedge B$	Comparison.Contrast
by contrast	$A \wedge B$	Comparison.Contrast
conversely	$A \wedge B$	Comparison.Contrast
Despite A, B. B, despite A.	$A \wedge B$	Comparison.Concession.Arg2-as-denier
A even as B. Even as B, A.	$A \wedge B$	Comparison.Concession.Arg1-as-denier Temporal.Synchronous Comparison.Concession.Arg1-as-denier
even before	$A \wedge B$	Temporal.Asynchronous.Precedence Comparison.Concession.Arg1-as-denier
even before then	$A \wedge B$	Temporal.Asynchronous.Succession Comparison.Concession.Arg2-as-denier
even if	$A \wedge B$	Comparison.Concession.Arg1-as-denier
even so	$A \wedge B$	Comparison.Concession.Arg2-as-denier
even then	$A \wedge B$	Temporal.Asynchronous.Precedence Comparison.Concession.Arg2-as-denier
even though	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier
even when	$A \wedge B$	Comparison.Concession.Arg1-as-denier Temporal.Asynchronous.Succession Comparison.Concession.Arg1-as-denier Temporal.Synchronous Comparison.Concession.Arg1-as-denier
even while	$A \wedge B$	Temporal.Synchronous Comparison.Concession.Arg1-as-denier
even with	$A \wedge B$	Comparison.Concession.Arg1-as-denier
finally	$A \wedge B$	Expansion.Conjunction
further	$A \wedge B$	Expansion.Conjunction
furthermore	$A \wedge B$	Expansion.Conjunction
A however B.	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier Comparison.Contrast Temporal.Synchronous Comparison.Contrast
in addition	$A \wedge B$	Expansion.Conjunction
in any case	$A \wedge B$	Comparison.Concession.Arg2-as-denier
in contrast	$A \wedge B$	Comparison.Contrast
in fact	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Conjunction
in the end	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Conjunction
in the meantime	$A \wedge B$	Temporal.Asynchronous.Succession Temporal.Synchronous Comparison.Contrast Temporal.Synchronous
in the meanwhile	$A \wedge B$	Temporal.Synchronous
indeed	$A \wedge B$	Comparison.Concession.Arg2-as-denier Expansion.Conjunction Expansion.Equivalence
like	$A \wedge B$	Comparison.Contrast Comparison.Similarity Expansion.Instantiation.Arg2-as-instance

likewise	$A \wedge B$	Expansion.Conjunction
meantime	$A \wedge B$	Temporal.Synchronous
meanwhile	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Conjunction Temporal.Synchronous Comparison.Concession.Arg2-as-denier Temporal.Synchronous Comparison.Contrast Temporal.Synchronous Comparison.Similarity Temporal.Synchronous
moreover	$A \wedge B$	Expansion.Conjunction
much less	$A \wedge B$	Expansion.Conjunction
neither A nor B.	$A \wedge B$	Comparison.Contrast Expansion.Conjunction
nevertheless	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast
no matter	$A \wedge B$	Comparison.Concession.Arg1-as-denier
nonetheless	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast
A; nor B.	$A \wedge B$	Comparison.Concession.Arg2-as-denier Expansion.Conjunction
not just A, but B.	$A \wedge B$	Comparison.Contrast Expansion.Conjunction
not just A, but also B.	$A \wedge B$	Comparison.Contrast Expansion.Conjunction
not only	$A \wedge B$	Comparison.Contrast Expansion.Conjunction
not only A, also B.	$A \wedge B$	Comparison.Contrast Expansion.Conjunction
not only A but B.	$A \wedge B$	Comparison.Concession.Arg2-as-denier Expansion.Conjunction
not only A but also B.	$A \wedge B$	Comparison.Contrast Expansion.Conjunction
on the contrary	$A \wedge B$	Comparison.Contrast
on the one hand A on the other B.	$A \wedge B$	Comparison.Contrast
on the one hand A on the other hand B.	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast
on the other hand	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast
only	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Exception.Arg2-as-excpt Expansion.Level-of-detail.Arg2-as-detail
A or B.	$A \wedge B$	Comparison.Concession+SpeechAct.Arg2-as-denier+SpeechAct Comparison.Concession.Arg2-as-denier Contingency.Condition+SpeechAct Contingency.Negative-condition.Arg1-as-negCond Expansion.Conjunction Expansion.Equivalence
plus	$A \wedge B$	Expansion.Conjunction
regardless	$A \wedge B$	Comparison.Concession.Arg2-as-denier
regardless of	$A \wedge B$	Comparison.Concession.Arg1-as-denier
separately	$A \wedge B$	Expansion.Conjunction Temporal.Synchronous Expansion.Conjunction
similarly	$A \wedge B$	Comparison.Similarity
simultaneously	$A \wedge B$	Temporal.Synchronous
still	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast Temporal.Asynchronous.Precedence Temporal.Synchronous
A then B.	$A \wedge B$	Expansion.Conjunction Temporal.Synchronous

though	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier Comparison.Contrast
A whatever B.	$A \wedge B$	Comparison.Concession.Arg1-as-denier
when	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier Comparison.Contrast Temporal.Synchronous Comparison.Contrast Temporal.Synchronous Contingency.Cause+Belief.Reason+Belief Temporal.Synchronous Contingency.Cause.Reason Temporal.Synchronous Contingency.Cause.Result Temporal.Synchronous Contingency.Condition+SpeechAct Temporal.Synchronous Contingency.Condition.Arg2-as-cond Temporal.Synchronous Expansion.Level-of-detail.Arg2-as-detail Temporal.Synchronous
whereas	$A \wedge B$	Comparison.Contrast
whether	$A \wedge B$	Comparison.Concession.Arg1-as-denier
while	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Concession.Arg2-as-denier Comparison.Contrast Comparison.Similarity Expansion.Conjunction Temporal.Synchronous Comparison.Concession.Arg1-as-denier Temporal.Synchronous Comparison.Concession.Arg2-as-denier Temporal.Synchronous Comparison.Contrast Temporal.Synchronous Expansion.Conjunction Temporal.Synchronous
with	$A \wedge B$	Comparison.Concession.Arg1-as-denier Comparison.Contrast Expansion.Conjunction Temporal.Synchronous
yet	$A \wedge B$	Comparison.Concession.Arg2-as-denier Comparison.Contrast Expansion.Conjunction

Table 21: Mapping from connectives to logical conjunction ($\wedge$), according to PDTB senses.

Connectives	**Binary Operator**	**Senses**
alternatively	$A \vee B$	Expansion.Disjunction Expansion.Substitution.Arg2-as-subst
as an alternative	$A \vee B$	Expansion.Disjunction
either A or B.	$A \vee B$	Contingency.Negative-condition.Arg1-as-negCond Expansion.Disjunction
A except B.	$A \vee B$	Expansion.Exception.Arg2-as-excpt
in other words	$A \vee B$	Expansion.Equivalence
instead	$A \vee B$	Expansion.Substitution.Arg2-as-subst
instead of	$A \vee B$	Expansion.Substitution.Arg1-as-subst
A, lest B. Lest B, A.	$A \vee B$	Contingency.Negative-condition.Arg1-as-negCond
not so much as	$A \vee B$	Expansion.Substitution.Arg2-as-subst
A, or B.	$A \vee B$	Expansion.Disjunction
or otherwise	$A \vee B$	Expansion.Disjunction
otherwise	$A \vee B$	Contingency.Negative-condition.Arg1-as-negCond Expansion.Exception.Arg1-as-excpt
rather	$A \vee B$	Expansion.Substitution.Arg2-as-subst
rather than	$A \vee B$	Expansion.Substitution.Arg1-as-subst
so much as	$A \vee B$	Expansion.Substitution.Arg2-as-subst
A unless B. Unless B, A.	$A \vee B$	Contingency.Negative-condition.Arg2-as-negCond
A, without B	$A \vee B$	Contingency.Negative-condition.Arg2-as-negCond

Table 22: Mapping from connectives to logical disjunction ($\vee$), according to PDTB senses.