MetaLogic: Logical Reasoning Explanations with Fine-Grained Structure

Yinya Huang\textsuperscript{1,2,*} Hongming Zhang\textsuperscript{2†} Ruixin Hong\textsuperscript{3} Xiaodan Liang\textsuperscript{1,4†} Changshui Zhang\textsuperscript{3} Dong Yu\textsuperscript{2}

\textsuperscript{1}Shenzhen Campus of Sun Yat-sen University \textsuperscript{2}Tencent AI Lab, Seattle \textsuperscript{3}Tsinghua University \textsuperscript{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.

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...
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 ¬, necessity □, and possibility ◊) and are joined by three binary operators on their logical relations (implication →, conjunction ∧, disjunction ∨). 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.
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.

Argumentation / Discourse Structures Previous works (Lawrence and Reed, 2019; Li et al., 2022) such as argumentation mining (Stub 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 \( S(0), S(1), \ldots, S(N) \) and atomic sentences \( p_k^{(n)} \subseteq S(n) \), according to which they generate the logic metagraph. The logic metagraph has three main components: (1) The meta structure \( G = (V, E) \), where \( E = ES \cup ER \), and \( ES \) and \( ER \) are the two meta edge types, support and rebut, respectively, between the meta nodes \( u^{(n)} \in V, n \leq N \). (2) The set of node formulae \( F \), where \( u^{(n)} := f_n \in F \). Each formula is joined by logical triples. \( f_n = \bigcap \{ m(p_i^{(n)}), m(p_j^{(n)}) \} \), where \( i \neq j, r \in \{ \rightarrow, \land, \lor \} \), and \( m \) is a combination in \( \{\neg, \square, \diamond \} \). (3) The set of degrees of certainty \( C \), defined by the combination format of \( \{\neg, \square, \diamond \} \).
Table 1: Senses, mutual definitions, reduction rules, and degrees of certainty of modal logic operators.

| Senses          | Classic | Morality   | Tense     | Belief               |
|-----------------|---------|------------|-----------|----------------------|
| □p              | The proposition \( p \) is necessary. | \( p \) is morally obligatory. | If it will always be the case that \( p \). | Things a person knows to be true. |
| ◇p              | The proposition \( p \) is possible. | \( p \) is morally permissible. | If it will sometimes be the case that \( p \). | Things that may be true as far as a person knows. |

4 The Logic Metagraph

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

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 (\( \rightarrow, \land, \lor \)). Multiple such logical triples are joined by conjunctions (\( \land \)). 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
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.

We recruit annotators from crowd-sourcing platforms. We first train annotators with a carefully designed annotation guideline and require them to pass an exam before the annotation to guarantee the annotation quality. For each passage, we invite two annotators. 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.

| Meta Structure | Meta Node |
|----------------|-----------|
| M-Node         | 57.80$^1$ |
| M-Edge         | 42.82$^1$ |
| L-Variable     | 65.46$^1$ |
| L-Relation     | 56.81$^1$ |

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 $\kappa$ over 60%. † indicates high agreement with $\kappa$ between 40% and 60%.

---

4Details are shown in Appendix B.

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

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

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

|                | Component 1 |        | Component 2 |        | Component 3 |        | Overall |        |
|----------------|-------------|--------|-------------|--------|-------------|--------|---------|--------|
|                | Node        | Step   | Formula     | Certainty |            |        |         |        |
|                | F1          | All    | F1          | All     | F1          | 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   |

|                | Component 1 |        | Component 2 |        | Component 3 |        | Overall |        |
|----------------|-------------|--------|-------------|--------|-------------|--------|---------|--------|
|                | Node        | Step   | Formula     | Certainty |            |        |         |        |
|                | F1          | All    | F1          | All     | F1          | All    | F1      | All    |
| 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 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 \( n_{reb\cap step\cap premise} = 36 \). 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).

|                | Passage Graph | Node | Form | Reb | Multi-step | Multi-premise |
|----------------|---------------|------|------|-----|------------|--------------|
| Avg.          | 1,000         | 1,000| 3,609| 1,500| 416        | 435          | 400        |
| Avg. Node     | 3.61          | 1.5  | 2.63 | 1.89 | 2.19       | 1.31         |
| Avg. Form     |               |      |      |      |            |              |
| Avg. Var      |               |      |      |      |            |              |
| Avg. Binary   |               |      |      |      |            |              |
| Avg. Global   |               |      |      |      |            |              |
| Avg. Local    |               |      |      |      |            |              |

4703
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.

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.

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.

| 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.
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.

6https://www.mindspore.cn/
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 | Question | Options |
|---------|----------|---------|
| 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. | The argument requires the assumption that | 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.
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.
Table 9: Question from ReClor with logical reasoning type: Weaken. Correct answer option in bold.

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

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.

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 meta-
graphs 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.

### 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., \(\text{sent1} \rightarrow \text{sent2}\)). 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

---

\(^7\)https://github.com/huggingface/transformers

---

**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.

---

**Task: Controller**

**Input:** CONTROL: proof: \(\text{sent1} \rightarrow \text{sent3}\); context: \(\text{sent1}: \text{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.} \ \text{sent2}: \text{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.} \ \text{sent3}: \text{therefore, in addition to the known planets, there must be at least one planet in our solar system that we have yet to discover.} \ \text{sent4}: \text{there is a belt of comets beyond the orbit of pluto with powerful gravitational pull.}

**Output:** \(\text{sent2} \Rightarrow \text{sent3}\)

**Task: Conclusion Module**

**Input:** CONCLUSION: \(\text{sent2}: \text{the dna of contemporary humans is significantly different from that of the neanderthal.} \ \text{sent3}: \text{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:** \(\text{sent3} \rightarrow \text{sent2}\)

**Task: Rebuttal Module**

**Input:** REBUTTAL: \(\text{sent1}: \text{recent unexpectedly heavy rainfalls in the metropolitan area have filled the reservoirs and streams; water rationing, therefore, will not be necessary this summer.} \ \text{sent2}: \text{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:** \(\text{sent2} \Rightarrow \text{sent1}\)

---

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, \(\neg v_1 \land \Diamond v_2\) is equivalent to \(\Diamond v_2 \land \neg v_1\), but \(\neg v_1 \rightarrow \Diamond v_2\) is not equivalent to \(\Diamond v_2 \rightarrow \neg v_1\). 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
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.
Table 14: Performance on two types of inference step. Per-operator F1 scores.

| Component 1 | Component 2 | Component 3 | Overall |
|------------|------------|-------------|---------|
| Node Step  | Formulae   | Certainty   |         |
| F1 All     | F1 All     | F1 All      |         |
| Acc All    | F1 All     | F1 All      |         |

| Support | Rebut | Operators |
|---------|-------|-----------|
| F1 All  | F1 All| → ∧ ∨ ¬  |
|         |       | 2 3       |
| N/A     |       | N/A       |

| Once (large) | 48.9 | 15.8 | 53.3 | 53.0 | 25.4 | 23.8 | 2.1 | 3.2 | 3.4 | 7.5 |
| 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 15: Detailed results on different ratios of MetaLogic training data. *: macro-F1.

| Ratio | Once (large) |-component1 |-component2 |-component3 | Overall |
|-------|--------------|------------|------------|------------|---------|
|       | F1 All       |            |            |            |         |
|       | F1 All       |            |            |            |         |
|       | Acc All      | F1 All     |            |            |         |
|       | F1 All       |            |            |            |         |

| Support | Rebut | Operators |
|---------|-------|-----------|
| F1 All  | F1 All| → ∧ ∨ ¬  |
|         |       | 2 3       |
| N/A     |       | N/A       |

| Ratio | Once (large) | F1 All | 47.8 | F1 All | 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 | 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 | 96.7 |
|       | 5%    | 20.4 | 1.5 | 26.3 | 26.0 | 5.5 | 1.8 | 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 | 53.4 |
|       | 1%    | 17.2 | 1.5 | 12.5 | 12.5 | 4.3 | 1.4 | 0.0 | 2.1 | 2.1 | 65.7 |

| Ratio | Multitask (large) | F1 All | 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 |

| Ratio | MetGen (large) | F1 All | 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.

4713
Table 17: Error cases for meta graph structure.

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; sent3 => sent1;

**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 fibrosis, 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;
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 accordingly | $A \rightarrow B$ | Contingency.Cause.Result                    |
| after             | $A \rightarrow B$ | Temporal.Asynchronous.Succession            |
| afterward         | $A \rightarrow B$ | Temporal.Asynchronous.Predence              |
| afterwards        | $A \rightarrow B$ | Temporal.Asynchronous.Predence              |
| B, as A.          | $A \rightarrow B$ | Contingency.Cause+Belief.Reason+Belief      |
| As A, B.          | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| as a result       | $A \rightarrow B$ | Contingency.Cause.Result                    |
| B, as if A.       | $A \rightarrow B$ | Comparison.Concession.Arg2-as-denier        |
| As if A, B.       | $A \rightarrow B$ | Comparison.Similarity                       |
| Because of A, B.  | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| B because of A.   | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| Because of A, B.  | $A \rightarrow B$ | Contingency.Condition+SpeechAct             |
| B, as long as A.  | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond          |
| As long as A, B.  | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond          |
| Before B, A.      | $A \rightarrow B$ | Temporal.Asynchronous.Predence              |
| A before B.       | $A \rightarrow B$ | Temporal.Asynchronous.Succession            |
| By A, B.          | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| B by A.           | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| by then           | $A \rightarrow B$ | Temporal.Asynchronous.Succession            |
| A consequently B. | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| B depending on A. | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond          |
| Depending on A, B.| $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond          |
| B, due to A.      | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| Due to A, B.      | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| B, earlier A, B.  | $A \rightarrow B$ | Temporal.Asynchronous.Succession            |
| A, finally, B.    | $A \rightarrow B$ | Temporal.Asynchronous.Predence              |
| B, for A.         | $A \rightarrow B$ | Comparison.Concession.Arg1-as-denier        |
| For A, B.         | $A \rightarrow B$ | Contingency.Cause.Reason                    |
| 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 |
|------------------|----------------|----------------------------------|
|                  |                | Contingency.Cause+Belief.Reason+Belief |
|                  |                | Contingency.Cause.Reason |
|                  |                | Contingency.Condition.Arg2-as-cond |
| B, from A.       | $A \rightarrow B$ | Contingency.Cause.Reason |
|                  |                | Expansion.Manner.Arg2-as-manner |
|                  |                | Contingency.Cause.Reason |
|                  |                | Expansion.Substitution.Arg1-as-subst |
| Given A, B.      | $A \rightarrow B$ | Contingency.Cause+Belief.Reason+Belief |
| B, given A.      | $A \rightarrow B$ | Contingency.Cause.Reason |
| A, hence B.      | $A \rightarrow B$ | Contingency.Cause.Result |
|                  |                | Comparison.Concession+SpeechAct. |
|                  |                | Arg2-as-denier+SpeechAct |
|                  |                | Comparison.Concession.Arg1-as-denier |
|                  |                | Comparison.Concession.Arg2-as-denier |
|                  |                | Comparison.Contrad |
|                  |                | 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 |
| If A, B.         | $A \rightarrow B$ | Contingency.Condition+SpeechAct |
| B, if A.         | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond |
|                  |                | Temporal.Synchronous |
|                  |                | Contingency.Condition.Arg2-as-cond |
|                  |                | 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.     | $A \rightarrow B$ | Contingency.Condition+SpeechAct |
|                  |                | Contingency.Condition.Arg2-as-cond |
|                  |                | Contingency.Purpose.Arg2-as-goal |
| If A then B.     | $A \rightarrow B$ | Contingency.Condition+SpeechAct |
|                  |                | Contingency.Condition.Arg2-as-cond |
|                  |                | Contingency.Purpose.Arg2-as-goal |
|                  |                | Expansion.Instantiation.Arg1-as-instance |
|                  |                | Expansion.Level-of-detail.Arg1-as-detail |
|                  |                | Expansion.Level-of-detail.Arg2-as-detail |
|                  |                | Expansion.Manner.Arg2-as-manner |
|                  |                | Temporal.Synchronous |
|                  |                | Contingency.Purpose.Arg2-as-goal |
|                  |                | Temporal.Synchronous |
|                  |                | Expansion.Level-of-detail.Arg1-as-detail |
|                  |                | Temporal.Synchronous |
|                  |                | B in A. |
|                  |                | Contingency.Purpose.Arg2-as-goal |
|                  |                | Expansion.Instantiation.Arg1-as-instance |
|                  |                | Expansion.Level-of-detail.Arg1-as-detail |
|                  |                | Expansion.Level-of-detail.Arg2-as-detail |
|                  |                | Expansion.Manner.Arg2-as-manner |
|                  |                | Temporal.Synchronous |
|                  |                | Contingency.Purpose.Arg2-as-goal |
|                  |                | Temporal.Synchronous |
|                  |                | Expansion.Level-of-detail.Arg1-as-detail |
|                  |                | Temporal.Synchronous |
| B in case A.     | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond |
|                  |                | Expansion.Instantiation.Arg2-as-instance |
|                  |                | Expansion.Level-of-detail.Arg1-as-detail |
|                  |                | Expansion.Level-of-detail.Arg2-as-detail |
| B, In fact, A.   | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond |
| B, A in fact.    | $A \rightarrow B$ | Contingency.Purpose.Arg2-as-goal |
|                  |                | 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.Purpose.Arg2-as-goal |
|                  |                | Expansion.Instantiation.Arg2-as-instance |
|                  |                | Expansion.Level-of-detail.Arg1-as-detail |
|                  |                | Expansion.Level-of-detail.Arg2-as-detail |
| B, in particular A. | $A \rightarrow B$ | Contingency.Purpose.Arg2-as-goal |
| 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 |
| A indeed B. | A → B |
| B, A indeed. | Contingency.Cause+Belief.Reason+Belief |
| B, insofar as A. | Contingency.Cause.Reason |
| Insofar as A, B. | Contingency.Cause.Result |
| A, B later. | Expansion.Conjunction |
| A, B later on. | Expansion.Instantiation.Arg2-as-instance |
| B, more accurately, A. | Expansion.Level-of-detail.Arg1-as-detail |
| A, next B. | Expansion.Level-of-detail.Arg2-as-detail |
| B, not only because of A. | Contingency.Cause.Reason |
| Now that A, B. | Contingency.Cause.Reason |
| B on A. | Contingency.Cause.Reason |
| Once A, B. | Contingency.Condition.Arg2-as-cond |
| B, once A. | Contingency.Cause.Reason |
| B only if A. | Contingency.Condition.Arg2-as-cond |
| B previously A. | Temporal.Asynchronous.Succession |
| B, since A. | Temporal.Asynchronous.Succession |
| Since A, B. | Contingency.Cause.Reason |
| B since before A. | Temporal.Asynchronous.Succession |
| A, so B. | Contingency.Condition.Arg2-as-cond |
| So as A, B. | Contingency.Condition.Arg2-as-cond |
| B so long as A. | Contingency.Condition.Arg2-as-cond |
| So long as A, B. | Contingency.Condition.Arg2-as-cond |
| A so that B. | Contingency.Condition.Arg2-as-cond |
| B, specifically, A. | Contingency.Condition.Arg2-as-cond |
| A subsequently B. | Contingency.Condition.Arg2-as-cond |
| B such as A. | Contingency.Condition.Arg2-as-cond |
| B, that is A. | Contingency.Condition.Arg2-as-cond |
| A then B. | Expansion.Conjunction |
| A thereafter B. | Expansion.Instantiation.Arg2-as-instance |
| A thereby B. | Expansion.Manner.Arg1-as-manner |
| A therefore B. | Expansion.Manner.Arg1-as-manner |
| A thus B. | Contingency.Negative-condition.Arg2-as-negCond |
| B till A. | Contingency.Negative-condition.Arg2-as-negCond |

4718
| A ultimately B. | $A \rightarrow B$ | Contingency.Cause.Reason
| | | Expansion.Conjunction
| | | Temporal.Asynchronous.Precedence
| | | Contingency.Cause.Result
| | | Temporal.Asynchronous.Precedence
| B untill A. | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond
| Untill A, B. | | Temporal.Asynchronous.Precedence
| | | Contingency.Condition.Arg2-as-cond
| | | Temporal.Asynchronous.Succession
| B upon A. | $A \rightarrow B$ | Temporal.Asynchronous.Succession
| Upon A, B. | | Contingency.Cause.Reason
| | | Temporal.Synchronous
| B, when A. | $A \rightarrow B$ | Contingency.Cause.Reason
| When A, B. | | Contingency.Condition.Arg2-as-cond
| | | Temporal.Asynchronous.Succession
| | | Contingency.Cause.Reason
| | | Temporal.Asynchronous.Succession
| | | Contingency.Condition.Arg2-as-cond
| | | Temporal.Asynchronous.Succession
| | | Contingency.Condition+SpeechAct
| | | Temporal.Asynchronous.Succession
| | | Contingency.Condition.Arg2-as-cond
| | | Temporal.Asynchronous.Succession
| B when and if A. | $A \rightarrow B$ | Temporal.Asynchronous.Succession
| | | Contingency.Condition.Arg2-as-cond
| B whenever A. | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond
| Whenever A, B. | | Contingency.Condition.Arg2-as-cond
| B, where A. | $A \rightarrow B$ | Contingency.Condition.Arg2-as-cond
| Where A, B. | | Contingency.Condition.Arg2-as-cond
| B, with A. | $A \rightarrow B$ | Contingency.Cause.Reason
| With A, B | | 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-as-manner
| B, without A. | $A \rightarrow B$ | Contingency.Cause.Reason
| Without A, B. | | Contingency.Cause.Reason
| | | Expansion.Level-of-detail.Arg2-as-detail
| | | Expansion.Manner.Arg2-as-manner

Table 20: Mapping from connectives to logical implication ($\rightarrow$), according to PDTB senses.
| Connectives | Binary Operator | Senses |
|-------------|----------------|--------|
| A, and B.   | $A \land B$   |        |
| additionally | $A \land B$   | Expansion.Conjunction |
| Albeit A, B. B, albeit A. | $A \land B$   | Comparison.Concession.Arg2-as-denier |
| along with | $A \land B$   | Expansion.Conjunction |
| also        | $A \land B$   | Expansion.Conjunction |
| although    | $A \land B$   | Expansion.Conjunction |
| as long as  | $A \land B$   | Contingency.Condition.Arg2-as-cond |
| as much as  | $A \land B$   | Expansion.Conjunction |
| as soon as  | $A \land B$   | Expansion.Conjunction |
| as though   | $A \land B$   | Expansion.Conjunction |
| as well     | $A \land B$   | Expansion.Conjunction |
| as well as  | $A \land B$   | Expansion.Conjunction |
| as          | $A \land B$   | Expansion.Conjunction |
| at the same time | $A \land B$   | Temporal.Synchronous |
| before and after | $A \land B$   | Temporal.Asynchronous.Precedence |
| besides     | $A \land B$   | Expansion.Conjunction |
| A, beyond B. Beyond B, A. | $A \land B$   | Expansion.Conjunction |
| both A and B. | $A \land B$   | Expansion.Conjunction |
| Expression                  | Arguments | Semantic Relations            |
|----------------------------|-----------|--------------------------------|
| but \( A \land B \)       |           | Comparison.Concession+SpeechAct.Arg2-as-denier |
|                           |           | Comparison.Concession.Arg2-as-denier |
|                           |           | Contingency.Cause+SpeechAct.Reason+SpeechAct |
|                           |           | Contingency.Cause.Reason |
|                           |           | Comparison.Concession.Arg2-as-denier |
|                           |           | Expansion.Conjunction |
|                           |           | Expansion.Exception.Arg2-as-excpt |
|                           |           | Temporal.Synchronous |
|                           |           | Comparison.Concession.Arg2-as-denier |
| A but also B.              | \( A \land B \) | Expansion.Conjunction |
| A but then B.              | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| A but then again B.        | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| by comparison              | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| by contrast                | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| conversely                 | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| Despite A, B, despite A.   | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| A even as B. Even as B, A. | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| even before                | \( A \land B \) | Temporal.Asynchronous.Precedence |
| even before then           | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| even if                    | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| even so                    | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| even then                  | \( A \land B \) | Temporal.Asynchronous.Precedence |
| even though                | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| even when                  | \( A \land B \) | Temporal.Asynchronous.Succession |
| even while                 | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| even with                  | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
| finally                    | \( A \land B \) | Expansion.Conjunction |
| furthermore                | \( A \land B \) | Expansion.Conjunction |
| A however B.               | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
|                           |           | Comparison.Concession.Arg2-as-denier |
|                           |           | Comparison.Concession.Arg2-as-denier |
|                           |           | Comparison.Contrast |
|                           |           | Temporal.Synchronous |
|                           |           | Comparison.Concession.Arg1-as-denier |
|                           |           | Expansion.Conjunction |
| in addition                | \( A \land B \) | Expansion.Conjunction |
| in any case                | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| in contrast                | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
| in fact                    | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
|                           |           | Comparison.Contrast |
|                           |           | Expansion.Conjunction |
| in the end                 | \( A \land B \) | Comparison.Concession.Arg1-as-denier |
|                           |           | Comparison.Concession.Arg2-as-denier |
|                           |           | Comparison.Contrast |
|                           |           | Expansion.Conjunction |
|                           |           | Temporal.Synchronous |
| in the meantime            | \( A \land B \) | Temporal.Synchronous |
|                           |           | Comparison.Contrast |
|                           |           | Temporal.Synchronous |
| in the meanwhile           | \( A \land B \) | Temporal.Synchronous |
| indeed                     | \( A \land B \) | Comparison.Concession.Arg2-as-denier |
|                           |           | Expansion.Conjunction |
|                           |           | Expansion.Equivalence |
| like                       | \( A \land B \) | Comparison.Contrast |
|                           |           | Comparison.Similarity |
|                           |           | Expansion.Instantiation.Arg2-as-instance |

4721
| Phrase                | Structure                  |
|----------------------|----------------------------|
| likewise             | $A \land B$ Expansion.Conjunction |
| meantime             | $A \land B$ Temporal.Synchronous |
| meanwhile            | $A \land B$ 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             | $A \land B$ Expansion.Conjunction |
| much less            | $A \land B$ Expansion.Conjunction |
| neither A nor B.     | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Comparison.Contrast |
| nonetheless          | $A \land B$ Comparison.Concession.Arg1-as-denier |
| no matter            | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Comparison.Contrast |
| nonetheless          | $A \land B$ Comparison.Concession.Arg1-as-denier |
| not just A, but also B.| $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Expansion.Conjunction |
| not only             | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Expansion.Conjunction |
| not only A, also B.  | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Expansion.Conjunction |
| not only A but B.    | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Expansion.Conjunction |
| on the contrary      | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Expansion.Conjunction |
| on the one hand A on the other B. | $A \land B$ Comparison.Concession.Arg2-as-denier |
| on the one hand A on the other hand B. | $A \land B$ Comparison.Concession.Arg2-as-denier |
| on the other hand    | $A \land B$ Comparison.Concession.Arg2-as-denier |
| only                 | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Expansion.Concession.Arg2-as-excpt |
|                      |                          | Expansion.Level-of-detail.Arg2-as-detail |
| A or B.              | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Contingency.Condition.Arg2-as-denier |
|                      |                          | Contingency.Negative-condition.Arg1-as-negCond |
|                      |                          | Expansion.Conjunction |
|                      |                          | Expansion.Equivalence |
| regardless           | $A \land B$ Comparison.Concession.Arg2-as-denier |
| regardless of        | $A \land B$ Comparison.Concession.Arg1-as-denier |
| separately           | $A \land B$ Temporal.Synchronous |
| similarly            | $A \land B$ Comparison.Similarity |
| simultaneously       | $A \land B$ Temporal.Synchronous |
| still                | $A \land B$ Comparison.Concession.Arg2-as-denier |
|                      |                          | Comparison.Contrast |
|                      |                          | Temporal.Aynchronous.Precendence |
|                      |                          | Temporal.Synchronous |
| A then B.            | $A \land B$ Expansion.Conjunction |
|                      |                          | Temporal.Synchronous |
| Connective | Logical Conjunction | Description |
|------------|---------------------|-------------|
| **though** | $A \land B$ | Comparison.Concession.Arg1-as-denier, Comparison.Concession.Arg2-as-denier, Comparison.Contrast |
| **A whatever B.** | $A \land B$ | Comparison.Concession.Arg1-as-denier |
| **when** | $A \land B$ | Comparison.Concession.Arg1-as-denier, Comparison.Concession.Arg2-as-denier, 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** | $A \land B$ | Comparison.Concession.Arg1-as-denier |
| **whether** | $A \land B$ | 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 |
| **while** | $A \land B$ | Comparison.Concession.Arg1-as-denier, Temporal.Synchronous, Comparison.Concession.Arg2-as-denier, Temporal.Synchronous, Comparison.Contrast, Temporal.Synchronous, Expansion.Conjunction, Temporal.Synchronous |
| **with** | $A \land B$ | Comparison.Concession.Arg1-as-denier, Comparison.Contrast, Expansion.Conjunction, Temporal.Synchronous |
| **yet** | $A \land B$ | Comparison.Concession.Arg2-as-denier, Comparison.Contrast, Expansion.Conjunction |

Table 21: Mapping from connectives to logical conjunction ($\land$), according to PDTB senses.
| Connectives      | Binary Operator | Senses                                      |
|------------------|-----------------|---------------------------------------------|
| alternatively    | $A \lor B$       | Expansion.Disjunction                       |
|                  |                  | Expansion.Substitution.Arg2-as-subst        |
| as an alternative | $A \lor B$       | Expansion.Disjunction                       |
| either A or B.   | $A \lor B$       | Contingency.Negative-condition.Arg1-as-negCond |
|                  |                  | Expansion.Disjunction                       |
| A except B.      | $A \lor B$       | Expansion.Exception.Arg2-as-excpt          |
| in other words   | $A \lor B$       | Expansion.Equivalence                       |
| instead          | $A \lor B$       | Expansion.Substitution.Arg2-as-subst        |
| instead of       | $A \lor B$       | Expansion.Substitution.Arg1-as-subst        |
| A, lest B.       | $A \lor B$       | Contingency.Negative-condition.Arg1-as-negCond |
| Lest B, A.       | $A \lor B$       | Expansion.Substitution.Arg2-as-subst        |
| not so much as   | $A \lor B$       | Expansion.Substitution.Arg2-as-subst        |
| A, or B.         | $A \lor B$       | Expansion.Disjunction                       |
| or otherwise     | $A \lor B$       | Expansion.Disjunction                       |
| otherwise        | $A \lor B$       | Contingency.Negative-condition.Arg1-as-negCond |
|                  |                  | Expansion.Exception.Arg1-as-excpt          |
| rather           | $A \lor B$       | Expansion.Substitution.Arg2-as-subst        |
| rather than      | $A \lor B$       | Expansion.Substitution.Arg1-as-subst        |
| so much as       | $A \lor B$       | Expansion.Substitution.Arg2-as-subst        |
| A unless B.      | $A \lor B$       | Contingency.Negative-condition.Arg2-as-negCond |
| Unless B, A.     | $A \lor B$       | Contingency.Negative-condition.Arg2-as-negCond |
| A, without B.    | $A \lor B$       | Contingency.Negative-condition.Arg2-as-negCond |

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