# A Question-Answer Driven Approach to Reveal Affirmative Interpretations from Verbal Negations

**Md Mosharaf Hossain,**[ϴ] **Luke Holman,**[ʊ] **Anusha Kakileti,**[ϴ] **Tiffany Iris Kao,**[ɤ]
**Nathan Raul Brito,**[ϴ] **Aaron Abraham Mathews,**[ϴ] and **Eduardo Blanco**[ʊ]
[ϴ]University of North Texas [ʊ]Arizona State University [ɤ]University of Texas at Austin
{mdmosharafhossain,nathanbrito,aaronmathews}@my.unt.edu lholman2@asu.edu
akakileti@gmail.com tiffanyiris1004@gmail.com eduardo.blanco@asu.edu

## Abstract

This paper explores a question-answer driven approach to reveal affirmative interpretations from verbal negations (i.e., when a negation cue grammatically modifies a verb). We create a new corpus consisting of 4,472 verbal negations and discover that 67.1% of them convey that an event actually occurred. Annotators generate and answer 7,277 questions for the 3,001 negations that convey an affirmative interpretation. We first cast the problem of revealing affirmative interpretations from negations as a natural language inference (NLI) classification task. Experimental results show that state-of-the-art transformers trained with existing NLI corpora are insufficient to reveal affirmative interpretations. We also observe, however, that fine-tuning brings small improvements. In addition to NLI classification, we also explore the more realistic task of generating affirmative interpretations directly from negations with the T5 transformer. We conclude that the generation task remains a challenge as T5 substantially underperforms humans.

## 1 Introduction

Negation can be understood as an operator that transforms the meaning of some expression into another expression whose meaning is in some way opposed to the original expression (Horn and Wansing, 2020). Typically, negated statements are less informative than affirmative statements (e.g., "Paris is not located in England" vs. "Paris is located in France"). Negated statements are also harder to process and understand by humans (Horn and Wansing, 2020). According to Horn (1989), negations carry affirmative meanings. These underlying affirmative meanings, which we refer to as *affirmative interpretations*, range from implicatures to entailments. For example, the negated statement (1) "Mary never drives long distances without a full tank of gas", carries at least the following affirmative interpretations: (1a) "Mary drives long distances," (1b) "Mary fills the gas tank before starting a long drive," and (1c) "Mary might drive short and medium distances without a full tank of gas."

| An extinct volcano is one that has not *erupted* in recent history. | |
|---|---|
| - Did something erupt? | *Yes* |
| - What erupted? | *An extinct volcano* |
| - When did something erupt? | *In the past* |
| Affirm. Intp: *An extinct volcano erupted in the past.* | |

| It was not *formed* by a natural process. | |
|---|---|
| - Was something formed? | *Yes* |
| - What was formed? | *It* |
| - What was something formed by? | *An artificial process* |
| Affirm. Intp: *It was formed by an artificial process.* | |

Table 1: Sentences containing negation, questions and answers about the affirmative counterpart of the main event, and the underlying affirmative interpretation.

In order to empower models to comprehend negation, most previous works target scope (Vincze et al., 2008; Morante and Daelemans, 2012) and focus (Blanco and Moldovan, 2011) detection (Section 2). Scope refers to the part of the meaning that is negated and focus refers to the part of the scope that is most prominently or explicitly negated (Huddleston and Pullum, 2002). Scope and focus detection plays a crucial role to understand what part of a negated statement is actually negated. These tasks do not, however, reveal affirmative interpretations—they tag tokens as belonging or not belonging to the scope and focus of a negation.

In this paper, we present a question-answer driven approach to reveal affirmative interpretations from verbal negations (i.e., when a negation cue grammatically modifies a verb). We adapt QA-SRL (He et al., 2015; FitzGerald et al., 2018) to collect questions and answers regarding the arguments of the affirmative counterpart of a negated predicate. Then, we manipulate the questions and answers to generate an affirmative interpretation. We find that generating and answering questions is



*Findings of the Association for Computational Linguistics: NAACL 2022*, pages 490 - 503

intuitive to non-experts (albeit they are native English speakers). Consider the examples in Table 1. Annotators first generate and answer a question regarding whether the main predicate in the sentence occurred (with unknown arguments at this point). Then, they generate and answer questions about the arguments of the affirmative counterpart of the main predicate. Arguments may come directly from the negated statement (e.g., What erupted? *An extinct volcano*) or using commonsense and world knowledge after reading the negated statement (e.g., When did something erupt? *In the past*). After collecting questions and answers, we automatically generate an affirmative interpretation in the form of a statement (e.g., *An extinct volcano erupted in the past*).

The main contributions of this paper are:[1]

1. A question-answer driven annotation schema to create AFIN, a corpus of verbal negations and their AFfirmative INterpretations (4,472 negations, 7,277 questions and answers, and 3,001 affirmative interpretations);
2. Corpus analysis indicating which predicate arguments are most often rephrased in the affirmative counterparts;
3. Casting the problem of revealing affirmative interpretations as a natural language inference task and showing that it is challenging for state-of-the-art transformers; and
4. Casting the problem of revealing affirmative interpretations as a generation task and showing that the T5 transformer substantially underperforms humans.

## 2 Related Work

Revealing affirmative interpretations from negations is a challenging endeavor. In the literature, researchers primarily seek to identify scope and focus of negation. The creation of the BioScope (Szarvas et al., 2008) and ConanDoyle-Neg (Morante and Daelemans, 2012) corpora spearheaded research on scope detection (Morante and Daelemans, 2009). Proposals include using traditional machine learning (Lapponi et al., 2012), off-the-shelf semantic parsers and semantic representations (Packard et al., 2014), and neural networks (Fancellu et al., 2016, 2017). PB-FOC (Blanco and Moldovan, 2011) is the largest corpus with focus of negation annotations. Recent proposals for focus detection include graph-based models with discourse information (Zou et al., 2014, 2015), neural networks with word-level and topic-level attention (Shen et al., 2019), and networks using scope information and context (Hossain et al., 2020). Scope and focus are useful to identify what is and what is not negated in a negated statement. Consider the second example in Table 1. Scope and focus do reveal that *It was formed*—everything but the focus (i.e., *by a natural process*) is affirmative—but provide no hints about *how* it was formed (i.e., *by an artificial process, artificially, etc)*. The main goal of this paper is to find these affirmative counterparts to generate affirmative interpretations.

More related to the work presented here, Sarabi et al. (2019) present a corpus of negations and their underlying affirmative interpretations (they call them positive interpretations). We are inspired by them but bypass several of their limitations. First, they only work with negations from Simple Wikipedia, a site devoted to English learners. As a result, their corpus uses (relatively) unsophisticated vocabulary and grammar. Second, they impose several restrictions on the negations they work with (e.g., negation cue modifies root verb, sentences between 6 and 25 tokens and not including certain tokens (because, until, etc.)). Third, their affirmative interpretations are restricted to a rephrasing of the statement containing negation with only one change: an argument of the negated predicate. In contrast, we barely impose restrictions on the negations we work with (no questions and no auxiliary verbs). More importantly, we introduce a question-driven approach that allow us to obtain multiple affirmative interpretations with increasing degrees of complexity (see examples in Table 3).

Recently, Jiang et al. (2021) study the problem of identifying commonsense implications of negations and contradictions. More specifically, they work with if-then rules such as *If X does not learn new things, then X does not gain new knowledge* and *If X does not leave the building, then X stays in the building*. These rules capture general commonsense knowledge about what happens if an event does not occur. Unlike them, we work with naturally occurring sentences that include negated predicate-argument structures with many arguments (agent, theme, manner, time, etc.). In addition, our affirmative interpretations reveal that predicates that are grammatically negated are actually factual (but with different arguments).

[1] Corpus and code available at https://github.com/mosharafhossain/AFIN.

| | Sent. | WH | AUX | SUB | VERB | OBJ1 | PREP | OBJ2 | |
|---|---|---|---|---|---|---|---|---|---|
| Predicate questions | (a) | | Was | something | formed | | | | ? |
| | (b) | | Does | something | kill | someone | | | ? |
| | (c) | | Will | someone | have | | to | do something | ? |
| Argument questions | (a) | What | was | something | formed | | by | | ? |
| | (b) | How often | does | something | kill | someone | | | ? |
| | (c) | Who | will | | have | | to | do something | ? |

Table 2: Predicate questions and argument questions (one per negated predicate) generated by annotators from the sentences (a) *It was not* formed *by a natural process.*, (b) *However, the ground shaking almost never* kills *people, [. . . ]*, and (c) *[. . . ], he hopes Australian teams will not* have *to travel so much to meet first class competition.*

## 3 A Question-Answer Driven Approach to Collect Affirmative Interpretations

This section outlines our approach to create AFIN, a corpus of verbal negations and their affirmative interpretations. We first describe the sources of naturally occurring negations in our corpus. Then, we outline the template-based approach to guide annotators in generating and answering questions about the *affirmative counterpart* of the negated predicate. Lastly, we describe the process to generate natural-language affirmative interpretations from the questions and answers.

### 3.1 Collecting Sentences Containing Negation

We start with the sentences in QA-SRL Bank 2.0 (FitzGerald et al., 2018), a corpus with 64,000 sentences across three domains: Wikipedia, Wikinews, and science textbooks (Kembhavi et al., 2017). Motivated by Fancellu et al. (2016), we select sentences containing negations checking for the following negation cues: *not*, *n't*, *no*, *never*, *without*, *nothing*, *none*, *nobody*, *nowhere*, and *neither* and *nor*. We only impose two restrictions: the sentences cannot be questions and the negation cues have to modify a verb that is not an auxiliary verb. We check the latter using universal dependencies as extracted by the parser in spaCy (Honnibal et al., 2020). We consider cues that directly or indirectly modify the verb, as exemplified in Figure 1. We will use *target* verb to refer to the negated verb in the remaining of the paper.

### 3.2 Generating and Answering Questions

Given a sentence and a target verb, our goal is to guide annotators to generate and answer questions about the (potential) affirmative counterpart of the target verb. First, they ask a *predicate question* to determine whether the affirmative counterpart of the target verb is factual (with unknown arguments). If it is, then they ask and answer *argument*

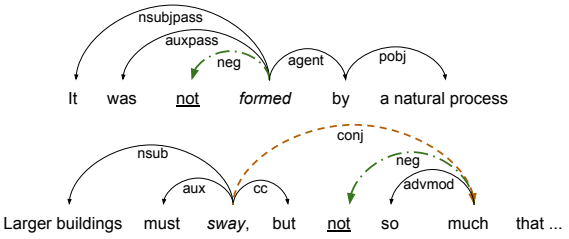


Figure 1: Illustration of the criteria to select negated verbs. We select all negations that modify non-auxiliary verbs either directly (top) or indirectly (bottom).

*questions* about the arguments of the affirmative counterpart of the target verb. Consider the following sentence: *However, no children resulted from the marriage*. The answer to the predicate question (Did anything result?) is *No*, thus no argument questions are considered. Now consider another sentence: *Cloning does not happen naturally*. The answer to the predicate question (Does something happen?) is *Yes*, thus annotators continue asking and answering argument questions: What happens? *Cloning* and How does it happen? *Artificially* (or *with human intervention*, for example).

**Template-Based Question Generation** In principle we could allow annotators to generate questions following their preferred wording. We found, however, that guiding them increases consistency and speed. To this end, we adapt the seven-slot template technique by He et al. (2015). For predicate questions (expected answer: Yes or No), we use the following combinations of slots: AUX x SUB x VERB x OBJ1 x PREP x OBJ2. For argument questions, we include an additional slot in the first position: WH. The full list of values for each slot are detailed in Appendix A. We provide below some examples for each slot.

- WH: Who, What, Whom, When, Where, etc.
- AUX: is, was, does, did, has, had, can, etc.
- SUB: something, or someone
- VERB: full conjugation of the target verb

- OBJ1: something, or someone
- PREP: by, to, for, with, about, of, or from
- OBJ2: someone, something, somewhere, do, doing, etc.

The templates allow annotators to generate a wide variety of questions. Table 2 shows several examples of predicate and argument questions generated from three target verbs. Note that humans are needed to choose values for each slot so that the resulting question is correct (right auxiliaries, conjugation, tense, number matching, etc.). Annotators generate questions in the following order of wh-words: *who* (or *what*) does/did (something) to *whom* (or *what*), *when*, *where*, *how*, *how much*, *how many*, *how long*, *how often*, and *why*. This order makes the generation of affirmative interpretations in natural language easier (Section 3.3).

**Answering Questions and Assigning Confidence Scores** Immediately after generating a question (i.e., before generating the next question), annotators answer it and indicate how confident they are in their answer. Note that several compatible answers are usually possible (e.g., *before* and *in the past* are usually interchangeable). Answers may come from the sentence containing the target verb and its arguments, or written by annotators using commonsense and world knowledge. Consider the following sentence: *The steep sides form because the lava cannot flow too far from the vent* (example (1) in Table 3). The answer to *What flows?* comes from the sentence: *Lava*. On the other hand, the answer to *Where does something flow?* is a rewrite of an argument of the target verb: *close to vents*. In the second example of Table 3, all answers come from the sentence with the target verb except *When was something classified?*, which is *In the past*.

Regardless of where answers come from, annotators assign a confidence score using a four-point Likert scale:

- 4: Extremely confident. I am certain that the answer is correct given the negated statement. For example, given "Scientists think that it will probably not *erupt* again," annotations answer *When did something erupt?* with *In the past* and assign a score of 4.
- 3: Very confident. My answer is very likely correct given the negated statement. For example, given "These volcanoes usually do not *produce* streams of lava," an annotator generated *How often does something produce?* and answered *Rarely* with a confidence score of 3.


Figure 2: Web interface to guide annotators in asking and answering questions. The screenshot shows the fillers for the VERB slot (i.e., the conjugation of *form*)

- 2: Moderately confident. My answer is likely correct given the negated statement. There are, however, many possible answers and my answer may be incorrect in an unlikely scenario. For example, given "It does not *release* carbon dioxide," an annotator assigned a confidence score of 2 to his answer to the question *What does something release? Fresh air*.
- 1: Slightly confident. My answer is probably correct, but there is no strong evidence in the sentence. For example, given "The second plot can not be *explained* using data," an annotator answered *How is something explained?* with *Using observations* and assigned a confidence score of 1. These answers often encode commonsense rather than an inference from the statement containing the target verb.

**Scaling the Annotation Process** Inspired by FitzGerald et al. (2018), we develop a web interface that facilitates the task of generating questions following our templates. More specifically, the interface auto-suggest to annotators the valid fillers for each slot. For example, if annotators start typing *W*, only fillers for the WH slot starting with *W* are suggested. The fillers for the next slot are suggested after the selection for the current slot is finalized. Figure 2 presents a screenshot of the interface with the auto-suggestions for the VERB slot (i.e., the conjugation of the target verb, *form*).

**Annotation Quality** Five undergraduate students who are native English speakers participated in the annotation process. They were trained in multiple sessions and conducted pilot annotations followed by discussion sessions before starting the annotations that resulted in the corpus described here. We do not calculate inter-annotator agree-

| | WH | AUX | SUB | VERB | PREP | | Answer | Affirmative Interpretation |
|---|---|---|---|---|---|---|---|---|
| (1) | What | | | flows | | ? | Lava | Lava flows. |
| | Where | does | something | flow | | ? | Close to vents | Lava flows close to vents. |
| (2) | What | was | | classified | | ? | Fungi | Fungi were classified. |
| | What | was | something | classified | as | ? | Plants | Fungi were classified as plants. |
| | When | was | something | classified | | ? | In the past | Fungi were classified as plants in the past. |

Table 3: Examples of questions and answers generated by annotators and the resulting affirmative interpretations. The sentences containing the negated predicates are (1) *The steep sides form because the lava cannot* flow *too far from the vent.* and (2) *Today, fungi are no longer* classified *as plants.* We do not show the OBJ1 and OBJ2 slots because they are empty for all the questions in these examples.

ment since two different answers to the same questions are likely to be correct. Consider the following sentence: "Scientists never *use* only one piece of evidence to form a conclusion." Two valid (and yet non-overlapping) answers to the question *What does someone use?* are *a reasonable amount of evidence* and *mathematical models*. The limitations of current automated metrics to determine whether these two answers are correct are well known (Liu et al., 2016), so we decided to conduct a manual evaluation. More specifically, we manually validated 479 questions and answers from a random sample of 200 target verbs in the corpus. A sixth person not involved in the generation and answering of the questions validated the 479 question-answer pairs as well as graded them with the same 4-point confidence scale. The validation phase revealed that (a) only 3% of the question-answer pairs are incorrect and (b) there is a strong correlation (Spearman: 0.71, Pearson: 0.70, p-value $< 0.005$ for both) between the scores.

### 3.3 Generating Affirmative Interpretations from Questions and Answers

We devise a rule-based approach in order to go from the questions generated and answered by annotators to an affirmative interpretation in natural language. Recall that annotators generate (and answer) questions in the following order: *who* (or *what*) does/did (something) to *whom* (or *what*), *when*, *where*, *how*, *how much*, *how many*, *how long*, *how often*, and *why*. Our approach is deterministic and manipulates answers depending on verb tense and number (which are obtained with part-of-speech tags and regular expressions).

We start with the answer to the first question (who (or what) does/did something?) in order to establish the subject of the affirmative counterpart of the target verb. Depending on whether the question uses the AUX slot, the affirmative interpretation also uses an auxiliary. Consider the examples in Table 3. In the first example, the question about the subject is *What flows?* and the answer is *Lava*, resulting in the initial affirmative interpretation *Lava flows*. Similarly, in the second example, the question is *What was classified?* and the answer is *Fungi*, resulting in the initial affirmative interpretation *Fungi were classified*.

Having generated an initial affirmative interpretation, the process continues adding arguments to the predicate-argument structure. We add them sequentially to the end of the affirmative interpretation in the order in which *argument questions* were generated and answered. Consider again the first example in Table 3. The only *argument question* left is *Where does something flow?*, which was answered with *Close to vents*. The initial affirmative interpretation becomes *Lava flows close to vents*. Since there are no additional questions, this is the final affirmative interpretation. Let us now consider the second example again. After incorporating the answer to the second question into the initial affirmative interpretation, we have *Fungi were classified as plants* (after including the preposition used in the question). Incorporating the answer to the third question, we have the final affirmative interpretation: *Fungi were classified as plants in the past*. The Appendix B provides additional details and special cases.

## 4 Corpus Analysis

The question-answer driven approach to generate and answer questions revealed that 3,001 out of the 4,472 (negated) target verbs carry an affirmative interpretation (67.1%). On average, annotators generated and answered 2.4 questions per target verb. Also, the average lengths of those questions and answers (in tokens) are 5.0 and 3.5, respectively. The average negated sentence is 25.8 tokens long, while its affirmative interpretation is 11.2 tokens

| min. confidence | 4 | $\geq 3$ | $\geq 2$ | $\geq 1$ |
|---|---|---|---|---|
| %verbs | 85.50 | 97.77 | 99.87 | 100.0 |

Table 4: Percentage of target verbs depending on the minimum confidence score assigned to any of the answers regarding the affirmative counterpart. Annotators almost always (97.77%) are *extremely* (4/4) or *very confident* (3/4) about their answers.

long, indicating that affirmative interpretations are much shorter than negated sentences. Appendix C provides additional details, including the distribution of wh-words in the questions.

Percentages, shown in Table 4, indicate that a vast majority of affirmative interpretations (85.5%) are generated from questions and answers about which annotators were *extremely confident* (confidence score: 4). The percentage raises to 97.77% if we include questions answered about which annotators were *very confident* (confidence score: 3).

Similar to Sarabi et al. (2019), we manually analyze 100 random examples from our corpus to find which arguments differ in the verb-argument structure of the (negated) target verb and the affirmative counterpart. We discovered these arguments primarily have the following functions (Frequencies and examples in Table 5):

- *Patient (or theme)* (24%). The most common argument is the person or thing that is affected or acted upon by the target verb. In the first example, we go from *workers had nothing* to *workers had only their labor*.
- *Manner* (23%). The second most common argument is the way in which the target verb takes place (the *how*). In the example, we go from *don't go through life with regrets* to *go through life with satisfaction*.
- *Quantity* (10%). Arguments expressing *specific* (e.g., four, three) or *abstract* quantities (e.g., many, less) represent 10% of changes in arguments. For example, we go from *Many mutations have no effect on the proteins* to *Some mutations have an effect on the proteins*.
- *Time* (10%). Tied in frequency with quantity, we observed arguments expressing temporal information. In the example, we go from *not allowed today* to *allowed in the past*.
- *Reason (or cause)* (9%). The fifth most common argument expresses the *why* of the target verb. We understand *why* widely, including reasons, causes, justifications, and explanations. In the example, we go from something not existing *without water* to *Earth has complexity and diversity because of water*.
- *Agent* (8%). The sixth most common argument is the person or thing who performs an event (i.e., the doer). In the example, we go from *an ideal capacitor not dissipating energy* to *a resistor dissipating energy*.
- *Other* (16%). Other functions (locations, purposes, recipients, etc.) account for 16% of arguments. Table 5 exemplifies a location change: from *cannot flow too far from the vent* to *flows close to the vents*.

## 5 Experiments and Discussion

AFIN consists of sentences containing verbal negations and their affirmative interpretations in natural language. We experiment casting the problem of obtaining affirmative interpretations from negation as a natural language inference task (Section 5.1) and as a generation task (Section 5.2).

### 5.1 Affirmative Interpretations and Natural Language Inference Classification

The sentences containing the (negated) target verb and the corresponding affirmative interpretations can be understood as the premises and hypotheses in a natural language inference (NLI) setting (Bowman et al., 2015). Very briefly, NLI is a classification task that determines whether a *premise* entails, is neutral with respect to, or contradicts a *hypothesis*. We label the premise-hypothesis pairs from AFIN as follows. If all the answers to questions used to generate an affirmative interpretation received the highest confidence score (4, Extremely confident), we label them *entailment* (85.5% of the target verbs). Otherwise (at least one answer received a confidence score between 1 and 3), we label them *neutral*. Note that contradiction examples cannot be derived from AFIN. Here we present two examples:

- Premise: *A dormant volcano no longer shows signs of activity.* Hypothesis: *A dormant volcano showed signs of activity in the past.* Premise *entails* hypothesis.
- Premise: *Respiratory infections such as pneumonia do not appear to increase the risk of COPD, at least in adults.* Hypothesis: *Respiratory infections appear to increase the risk of COPD in elderly.* Premise is *neutral* with respect to the hypothesis.

| Category | % | Example |
|---|---|---|
| Patient | 24 | Many workers, who *had* nothing but their labour to sell, became factory workers out of necessity. → *Many workers had only their labour to sell.* |
| Manner | 23 | I don't *go* through life with regrets. → *I go through life with satisfaction.* |
| Quantity | 10 | Many mutations *have* no effect on the proteins they encode. → *Some mutations have an effect on the proteins they encode.* |
| Time | 10 | The use of asbestos is not *allowed* today. → *The use of asbestos was allowed in the past.* |
| Reason | 9 | Without water, life might not be able to exist on Earth and it certainly would not *have* the tremendous complexity and diversity that we see. → *Earth has complexity and diversity because of water.* |
| Agent | 8 | Unlike a resistor, an ideal capacitor does not *dissipate* energy. → *A resistor dissipates energy.* |
| Others | 16 | The steep sides form because the lava can not *flow* too far from the vent. → *Lava flows close to vents.* |

Table 5: Analysis of the arguments that differ in the target verb and the corresponding affirmative counterpart. Categories refer to the function in the verb-argument structure. A wavy underline indicates the new argument in the affirmative counterpart.

| | | RoBERTa | | | XLNet | | |
|---|---|---|---|---|---|---|---|
| | Tested w/ | P | R | F1 | P | R | F1 |
| MNLI | MNLI-dev | 88 | 88 | 88 | 87 | 87 | 87 |
| | MNLI-dev* | 92 | 87 | 89 | 91 | 85 | 88 |
| | AFIN | 55 | 43 | 48 | 54 | 42 | 47 |
| SNLI | SNLI-dev | 92 | 92 | 92 | 91 | 91 | 91 |
| | SNLI-dev* | 93 | 90 | 92 | 93 | 90 | 92 |
| | AFIN | 56 | 37 | 45 | 57 | 38 | 46 |
| RTE | RTE-dev | 76 | 76 | 76 | 70 | 68 | 69 |
| | AFIN | 52 | 53 | 52 | 53 | 55 | 54 |

Table 6: Precision, Recall, and F1 scores (macro average) obtained with RoBERTa and XLNet trained with MNLI, SNLI, and RTE. We provide results with the original development set in each benchmark, the subsets that only contain *entailment* and *neutral* pairs (*), and the premise-hypothesis pairs derived from AFIN, our corpus. Transformers trained with any of the benchmarks perform substantially worse with AFIN.

| | RoBERTa | | | XLNet | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| MNLI-training | 55 | 43 | 48 | 54 | 40 | 46 |
| + 70% of AFIN | 72 | 51 | 60 | 61 | 51 | 55 |
| SNLI-training | 58 | 36 | 45 | 60 | 38 | 47 |
| + 70% of AFIN | 42 | 50 | 46 | 61 | 52 | 56 |
| RTE-training | 51 | 52 | 52 | 52 | 54 | 53 |
| + 70% of AFIN | 56 | 53 | 54 | 61 | 55 | 58 |

Table 7: Results obtained training with (a) MNLI, SNLI, or RTE and (b) 70% of AFIN, and evaluating with 30% of AFIN. Fine-tuning improves results, but transformers substantially underperform the original development splits (see Table 6).

**Transformers and Existing NLI Benchmarks** At first, we seek to investigate whether state-of-the-art transformers trained with existing NLI benchmark can solve the premise-hypothesis pairs derived from AFIN. Note that to do so, they would need to make inference in the presence of negation. We experiment with (a) two transformers: RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019), and (b) three NLI benchmarks: MNLI (Williams et al., 2018), SNLI (Bowman et al., 2015), and RTE (part of the GLUE benchmark (Wang et al., 2018)). We fine-tuned the transformers with the training split of each benchmark and conduct three evaluations: with (a) the development split of each benchmark, (b) the subsets of (a) that only contain *entailment* and *neutral* pairs, and (c) all premise-hypothesis pairs derived from AFIN. Note that neither RTE nor AFIN have pairs annotated *contradiction*. Appendix D.1 details the training procedure.

Table 6 presents the results. While both transformers obtain roughly the same results when evaluated with the three labels or only *entailment* and *neutral* pairs, we observe substantial drops in F1 score when evaluated with AFIN, around 46% with MNLI and 51% with SNLI.

We observe a similar pattern with RTE, although the drop is relatively small with XLNet (note, however, that XLNet does much worse than RoBERTa (69 vs. 76), whose performance drops 32%). We hypothesize that RTE obtains better results because it does not contain *contradiction* pairs. These results show that current benchmarks are not enough to identify inferences between a negation and its affirmative interpretation.

**Fine-tuning with AFIN** The next experiments examine whether fine-tuning helps transformers identify inference relations in the premise-hypothesis pairs generated from AFIN. To do so, we fine-tune the transformers not only with an ex-

| | BLEU-2 | chrf++ | METEOR |
|---|---|---|---|
| Negated sent. | 26.5 | 50.5 | 43.5 |
| + target verb | 33.6 | 57.3 | 51.9 |

Table 8: Evaluation results obtained with BLEU-2, chrf++, and METEOR between human and system generated affirmative interpretations.

isting benchmark (MNLI, SNLI, or RTE), but also 70% of the pairs derived from AFIN. Then, we evaluate with 30% of the pairs derived from AFIN.

Table 7 presents the results. Perhaps unsurprisingly, fine-tuning with AFIN allows the transformers to correctly identify few more *entailment* and *neutral* pairs (F1 scores: 45–53 vs. 46–60). We note, however, that no matter how we combine transformers and NLI benchmarks, the results are substantially below those obtained with the original development split (F1 scores: 46–60 vs. 69–92).

### 5.2 Generating Affirmative Interpretations

Casting the problem as a natural language inference task is worthwhile but unrealistic: the affirmative interpretations to be verified (are they entailed by the sentence with the negation?) are not readily available. In our next experiments, we investigate a realistic formulation of the problem: *generate* affirmative interpretations given a sentence with a negation. In order to do so, we split AFIN as follows: 70% for training, 15% for development, and the remaining 15% for test.

**Experimental Setup** We perform the experiments with the T5-Large transformer (Raffel et al., 2020), which can generate text through a supervised learning setup. In particular, we train T5 to generate affirmative interpretations using two inputs: (a) only the sentence containing the (negated) target verb (i.e., the negated sentence), and (b) the negated sentence concatenated with the target verb. The second setup investigates whether inputting the target verb with the negated sentence aids in generating affirmative interpretation about that target verb. Additional details on the training procedure are provided in Appendix D.2.

**Results and Analysis** After the training process with both setups, we obtain evaluation scores using three automatic metrics: BLEU-2 (Papineni et al., 2002), chrf++ (Popović, 2017), and METEOR (Banerjee and Lavie, 2005). We calculate these metrics comparing the human- and T5-generated affirmative interpretations from the test split (Table 8). Evidently, the system provided with the target verb shows comparatively better scores than the system without the target verb (e.g., BLEU-2: 33.6 vs 26.5). Based on the scores from the best setting, T5 achieves some capability to automatically generate affirmative interpretations.

| | Confidence Scores | | | | |
|---|---|---|---|---|---|
| | 4 | 3 | 2 | 1 | 0 |
| AFIN (upper bound) | 86.2 | 11.6 | 2.0 | 0.2 | n/a |
| T5-Large | | | | | |
| Negated sent. | 32.0 | 15.3 | 12.0 | 3.3 | 37.3 |
| + target verb | 43.3 | 10.0 | 15.3 | 4.0 | 27.3 |

Table 9: Percentage of affirmative interpretations assigned each confidence score in (a) the AFIN test set and (b) those generated by T5 (not providing and providing the target verb). T5 substantially underperforms AFIN, which is a human upper bound.

While useful, automatic metrics only provide a partial picture about the quality of affirmative interpretations, as outlined in Section 3.2. Therefore, we manually evaluate the affirmative interpretations generated by T5. In particular, the same annotator that validated a sample of AFIN validated the output of T5 with the confidence scores provided in Section 3.2.[2] Note that this time we added a new score of 0 to indicate that an affirmative interpretation is incorrect.

Table 9 provides the results. The scores assigned to AFIN represent an upper bound. We observe that explicitly providing the (negated) target verb is beneficial as it allows T5 to generate many more *extremely confident* affirmative interpretations (32% vs. 43.3%). We observe, however, that T5 faces challenges generating affirmative interpretations. First, over a quarter (27.3%) are incorrect. Second, compared to AFIN (i.e., human annotators), T5 only generates about half (43.3% vs. 86.2%) of affirmative interpretations that an evaluator is *extremely confident* about (confidence score: 4).

**Qualitative Analysis** In addition to confidence scores, we also analyze when T5 faces the biggest challenges generating affirmative interpretations. To this end, we randomly selected 150 instances from the test split. Then, we manually annotated the functions of the arguments that should be replaced in the affirmative interpretations with the same categories than the ones discussed in Section 4. We present the confidence score analysis in Fig-

[2]The only difference is that the affirmative interpretations come from T5 instead of a human annotator.

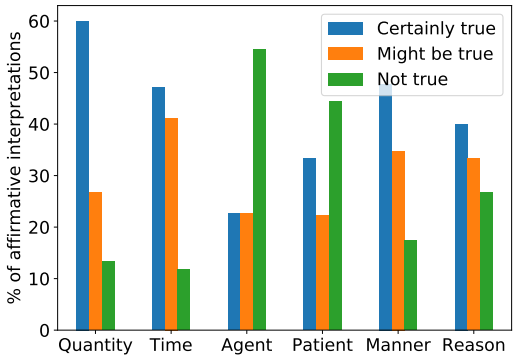


Figure 3: Analysis of scores assigned to the affirmative interpretations generated by T5. Scores are much lower when the argument that has to be changed to generate the affirmative interpretation is an *agent* or *patient*.

ure 3. For convenience, we show scores in three groups: *certainly true* (score: 4), *might be true* (scores from 1 to 3), and *not true* (score: 0).

We observe that it is comparatively easy for T5 to generate *certainly true* affirmative interpretations when the argument to be replaced contains a quantity (*certainly true* vs. *not true*: 60% vs 13.3%). Therefore, T5 learned some patterns to replace quantities in the affirmative interpretations. For example, from negation "Schools can not *charge* students more than US$5 to defray the cost of insurance," T5 correctly generates "Schools can charge students US$5 to cover the cost of insurance." Despite the relatively success with quantities, less than 50% of all affirmative interpretations that require replacing an argument in any other category are deemed certainly correct. *Agent* and *patient* are the categories T5 finds most challenging—these affirmative interpretations are more often deemed *not true* than *certainly true*. T5 often generates affirmative interpretations in these categories by deleting the negation cue and fixing verb tense and auxiliaries to form a grammatical—but incorrect—affirmative interpretation. For example, given "Ryanair have also sacked veteran pilot John Goss for appearing on the show, the only pilot interviewed who did not *seek* anonymity," T5 generates "Veteran pilot John Gosson sought anonymity."

## 6 Conclusions

We have proposed a question-answer driven approach to reveal affirmative interpretations from verbal negations. Annotators generate and answer questions regarding the affirmative counterpart of a negated verb, and then we generate from them an affirmative counterpart in natural language. Through analyses, we have shown that 67.1% of verbal negations convey that the negated event is actually factual. More importantly, we observe many categories in the arguments that are replaced in the affirmative interpretations (patient, manner, quantity, time, reason, etc.). The experiments show that transformers struggle substantially when we cast the problem as NLI. Doing so, however, is an unrealistic scenario: affirmative interpretations are not readily available to be fed into a natural language inference classifier. Further, we observe very limited success generating affirmative interpretations given as input a sentence containing verbal negation. We argue that generating affirmative interpretation is the realistic scenario and propose doing so as a challenging generation task requiring a combination of language comprehension, commonsense, and world knowledge currently out of reach for state-of-the-art models.

## Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1845757. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. The Titan Xp used for this research was donated by the NVIDIA Corporation. Computational resources were also provided by the UNT office of High-Performance Computing. Additionally, we utilized computational resources from the Chameleon platform (Keahey et al., 2020). We also thank the anonymous reviewers for their insightful comments.

## References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Eduardo Blanco and Dan Moldovan. 2011. Semantic representation of negation using focus detection. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 581–589, Portland, Oregon, USA. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large anno-

tated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 495–504, Berlin, Germany. Association for Computational Linguistics.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Nicholas FitzGerald, Julian Michael, Luheng He, and Luke Zettlemoyer. 2018. Large-scale QA-SRL parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2051–2060, Melbourne, Australia. Association for Computational Linguistics.

Luheng He, Mike Lewis, and Luke Zettlemoyer. 2015. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 643–653, Lisbon, Portugal. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python, 2020. *URL https://doi. org/10.5281/zenodo*, 1212303(6).

Laurence Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Laurence R. Horn and Heinrich Wansing. 2020. Negation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2020 edition. Metaphysics Research Lab, Stanford University.

Md Mosharaf Hossain, Kathleen Hamilton, Alexis Palmer, and Eduardo Blanco. 2020. Predicting the focus of negation: Model and error analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8389–8401, Online. Association for Computational Linguistics.

Rodney D. Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge University Press.

Liwei Jiang, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. 2021. "I'm not mad": Commonsense implications of negation and contradiction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4380–4397, Online. Association for Computational Linguistics.

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S. Gunawi, Cody Hammock, Joe Mambretti, Alexander Barnes, François Halbach, Alex Rocha, and Joe Stubbs. 2020. Lessons learned from the chameleon testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pages 4999–5007.

Emanuele Lapponi, Erik Velldal, Lilja Øvrelid, and Jonathon Read. 2012. UiO 2: Sequence-labeling negation using dependency features. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 319–327, Montréal, Canada. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Roser Morante and Walter Daelemans. 2009. A metalearning approach to processing the scope of negation. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 21–29, Boulder, Colorado. Association for Computational Linguistics.

Roser Morante and Walter Daelemans. 2012. ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. European Language Resources Association (ELRA).

Woodley Packard, Emily M. Bender, Jonathon Read, Stephan Oepen, and Rebecca Dridan. 2014. Simple negation scope resolution through deep parsing: A semantic solution to a semantic problem. In *Proceedings of the 52nd Annual Meeting of the Association*

*for Computational Linguistics (Volume 1: Long Papers)*, pages 69–78, Baltimore, Maryland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

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):1–67.

Zahra Sarabi, Erin Killian, Eduardo Blanco, and Alexis Palmer. 2019. A corpus of negations and their underlying positive interpretations. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 158–167, Minneapolis, Minnesota. Association for Computational Linguistics.

Longxiang Shen, Bowei Zou, Yu Hong, Guodong Zhou, Qiaoming Zhu, and AiTi Aw. 2019. Negative focus detection via contextual attention mechanism. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2251–2261, Hong Kong, China. Association for Computational Linguistics.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: Annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, BioNLP '08, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.

Veronika Vincze, György Szarvas, Richárd Farkas, György Móra, and János Csirik. 2008. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC bioinformatics*, 9(11):S9.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2014. Negation focus identification with contextual discourse information. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 522–530, Baltimore, Maryland. Association for Computational Linguistics.

Bowei Zou, Guodong Zhou, and Qiaoming Zhu. 2015. Unsupervised negation focus identification with word-topic graph model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1632–1636, Lisbon, Portugal. Association for Computational Linguistics.

## A Additional Details on Template-based Question Generation

This section provides additional details for the slots in the template-based question generation presented in Section 3.2 of the paper.

- WH indicates wh-words to generate the argument questions. The complete set of options we use are as follows: *who, what, whom, when, where, how, how much, how many, how long, how often*, and *why*.
- AUX indicates auxiliary verbs. The predicate questions always start with an auxiliary. However, argument questions may or may not contain an auxiliary verb (See examples in Table 3 of the paper). We avail the below list of auxiliary verbs for annotators: *is, was, does, did, has, had, can, could, may, might, will, would, should*, and *must*.
- SUB refers to subjects. Similar to He et al. (2015), we only avail *someone* or *something*, indicating placeholder for the subject position.
- VERB indicates the full conjugation of the target verb.

- OBJ1 refers to the options for objects. Similar to SUB, we only avail *someone* or *something*, indicating placeholders for objects.
- PREP refers to prepositions. We avail a short list of common prepositions: *by, to, for, with, about, of*, and *from*.
- OBJ2 refers to the additional options for objects. The complete list includes the following: *someone, something, somewhere, do, doing, do something*, and *doing something*.

## B Additional Details on Generating Affirmative Interpretations from Questions and Answers

The process to generate affirmative interpretations from questions and answers is robust but not foolproof from a grammatical standpoint. Note that the semantics of the affirmative interpretation is dictated by the questions and answers, and our evaluation determined that only 3% are incorrect (Section 3.2). We manually validated the final affirmative interpretations for grammaticality and found that 9% have errors. For example, consider *Most plastics do not form crystals.* The questions and answers are as follows: *What forms something? Plastic*, *What does something form? Crystals*, and *How many form? Few*.[3] These result in the affirmative interpretation *Plastic forms crystals few*, which places *few* incorrectly. We manually fix all the grammatical issues we found in the affirmative interpretations. Table 10 provides additional examples (Similar to Table 3 in the paper).

## C Additional Details on Corpus Analysis

Table 11 presents percentages of negated sentences and their affirmative interpretations in several length buckets. In the corpus, sentences with negation are fairly long, for example, 29.76% of them are longer than 29 tokens. The affirmative interpretations, however, are much shorter, with 79.9% being under 15 tokens.

In Table 12, we report the percentage of the argument questions that start with each wh-word (first column) and the percentage of negated verbs that contain a wh-word (second column). For example, 52.61% of all the argument questions start with the wh-word *what*, and 92.27% of all the negated verbs contains at least one question that starts with *what*.

[3]An alternative could be answering *What forms something?* with *Few plastics* (and skip the question starting with *How many*).

## D Training Procedure and Hyperparameters

### D.1 Affirmative Interpretations and Natural Language Inference Classification

For all the experiments mentioned in Section 5.1 in the paper, we use Huggingface implementation (Wolf et al., 2019) of the transformer systems. In addition, we utilize the base architecture (12-layer, 768-hidden, 12-heads) of transformers and their pretrained weights. We accept the default setting for most of the hyperparameters, except a few carefully selected to fine-tune the systems. Table 13 shows the hyperparameters used to fine-tune RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) on the three NLI corpora. Our code is available at https://github.com/mosharafhossain/AFIN.

### D.2 Generating Affirmative Interpretations

In order to generate affirmative interpretations for both input configurations (Section 5.2 in paper), we use the same set of hyperparameters discovered through cross-validation to tune the T5-Large system. Further, for the setup that adds the target verb with the sentence containing negation, we use two prefixes[4] (one for the *target verb* and another for the *negated sentence*) to create a single text before encoding it and passing to the T5 system. During the training process, we stop as soon as the loss (T5 uses *cross-entropy*) in the development split does not increase for 10 epochs. Thus, the final model is the one that produces the lowest loss in the development split. Table 14 provides the list of hyperparameters values in our experiments. In each run, the model requires approximately three hours to train on a single NVIDIA Tesla K80 GPU. The code is available at https://github.com/mosharafhossain/AFIN.

[4]https://huggingface.co/docs/transformers/model_doc/t5

| | WH | AUX | SUB | VERB | PREP | | Answer | Affirmative Interpretation |
|---|---|---|---|---|---|---|---|---|
| (1) | What | | | happens | | ? | Reflection | Reflection happens. |
| | What | does | something | happen | with | ? | Any type of waves | Reflection happens with any type of waves. |
| (2) | What | was | | made | by | ? | It | It was made. |
| | What | was | something | made | by | ? | Inanimate organisms | It was made by inanimate organisms. |
| (3) | What | has | something | | | ? | Later life forms | Later life forms have. |
| | What | does | something | have | | ? | The ability to photosynthesize | Later life forms have the ability to photosynthesize. |
| (4) | What | | | rises | | ? | The Sun | The Sun rises. |
| | When | does | something | rise | | ? | In all seasons | The Sun rises in all seasons. |
| | Where | does | something | rise | | ? | In the sky | The Sun rises in all seasons in the sky. |
| | How much | does | something | rise | | ? | Very low | The Sun rises in all seasons in the sky very low. |
| (5) | Who | | | returned | | ? | Locke | Locke returned. |
| | When | did | someone | return | | ? | After the Glorious Revolution | Locke returned after the Glorious Revolution. |
| | Where | did | someone | return | | ? | Home | Locke returned after the Glorious Revolution home. |

Table 10: Examples of questions and answers generated by annotators and the resulting affirmative interpretations. The sentences containing the negated predicates are (1) *Reflection can* happen *with any type of waves, not just sound waves*, (2) *It was not* made *by living organisms*, (3) *The earliest life forms did not* have *the ability to photosynthesize*, (4) *Even in summer, the Sun never* rises *very high in the sky*, and (5) *Locke did not* return *home until after the Glorious Revolution*. We do not show the OBJ1 and OBJ2 slots because they are empty for the questions in these examples.

| Lengths | %Neg. Sentences | %Affirm. Interpretations |
|---|---|---|
| <10 | 4.10 | 43.12 |
| 10–14 | 13.93 | 36.79 |
| 15–19 | 21.09 | 13.53 |
| 20–24 | 19.06 | 4.30 |
| 25–29 | 12.06 | 1.53 |
| >29 | 29.76 | 0.73 |
| All | 100 | 100 |

Table 11: Percentages of negated sentences and affirmative interpretations in different length buckets. Length is measured in tokens. The average length of a negated sentence and its affirmative interpretation is 25.8 and 11.2, respectively.

| | % | %verb with |
|---|---|---|
| What | 52.61 | 92.27 |
| Who | 17.27 | 39.19 |
| when | 9.89 | 23.89 |
| how | 7.09 | 17.09 |
| where | 6.31 | 15.19 |
| why | 3.60 | 8.73 |
| how much | 1.65 | 4.00 |
| how often | 0.63 | 1.53 |
| how many | 0.51 | 1.23 |
| how long | 0.41 | 1.00 |
| whom | 0.03 | 0.07 |

Table 12: Percentages of argument questions starting with each wh-word and percentages of negated verbs containing questions that start with each wh-word.

| Hyperparameter | RTE | | SNLI | | MNLI | |
|---|---|---|---|---|---|---|
| | RoBERTa | XLNet | RoBERTa | XLNet | RoBERTa | XLNet |
| Batch size | 16 | 8 | 32 | 32 | 32 | 32 |
| Learning rate | 2e-5 | 2e-5 | 1e-5 | 1e-5 | 2e-5 | 2e-5 |
| Epochs | 10 | 50 | 3 | 3 | 3 | 3 |
| Weight decay | 0.0 | 0.0 | 0.1 | 0.1 | 0.0 | 0.0 |

Table 13: Hyperparameters for finetuning the transformer systems used in Section 5.1 in the paper.

| Hyperparameter | |
|---|---|
| Max Epochs | 50 |
| Batch Size | 4 |
| Sentence max length | 128 |
| Optimizer | Adafactor |
| Learning rate | 1e-5 |
| Weight decay | 5e-6 |
| Warmup epoch | 5 |
| Accumulate step | 1 |
| Grad_clipping | 5.0 |
| Top_k | 50 |
| Top_p | 0.95 |
| Repetition_penalty | 2.5 |

Table 14: Hyperparameters for finetuning T5-Large on AFIN (Section 5.2 in paper).