Quantifier Scope Disambiguation

KC Tsiolis

Summer 2020

1 Background

Quantifier scope disambiguation (QSD) is a hard problem in natural language processing. In fact, recent literature on the problem is scarce (at least from an NLP perspective), and no neural network approach has been attempted to solve the problem, which stands in stark contrast to many other NLP tasks.

Quantifier scope ambiguity arises in sentences with multiple quantifiers. It is best illustrated with an example (Srinivasan and Yates, 2009):

- A doctor lives in every city.

We have underlined the two quantifiers in the sentence: “A” and “every”. The former is an existential quantifier and the latter is a universal quantifier. There are two possible interpretations of this sentence. The first is that there exists a doctor who resides in every city. In this case, we would say that the existential quantifier “a” takes **wide scope**. Consequently, the universal quantifier “every” takes **narrow** scope. This reading, where the first quantifier takes wide scope, is referred to as the **surface scope reading**. The second and more plausible interpretation of the example sentence is that for every city, there exists a doctor who lives there (not necessarily the same doctor). In this case, “every” takes wide scope. This reading, where the last quantifier takes wide scope, is called the **inverse scope reading**. For human readers, it is clear that the second reading is preferred. But what about automatic systems?

Higgins and Sadock (2003) present the first machine learning approach for quantifier scope disambiguation. They construct a dataset of 890 sentences from the Penn Treebank Wall Street Journal corpus (Marcus et al., 1993). Each sentence contains two quantifiers, where either the first takes wide scope, the second takes wide scope, or neither outscopes the either. The latter is the case for the majority of the sentences in the dataset. Their approach relies on making use of syntactic features from the parse trees that come with the PTB sentences. They experiment with three different models: Naive Bayes, maximum-entropy, and a single-layer perceptron, which all perform similarly on the task and outperform the most frequent label baseline.

Andrew and MacCartney (2004) present a dataset of their own, extracting sentences from GRE and LSAT logic games. Once again, each sentence contains two quantifiers. They have four class labels: FIRST (surface scope reading), SECOND (inverse scope reading), EQUIVALENT (surface and inverse scope readings are equivalent) AND INDEPENDENT (no scopal interaction). They make use of lexical features from the quantifiers themselves, as well as syntactic parse features. They apply Naive Bayes, logistic regression, and support vector machines to their dataset. When focusing solely on the FIRST and SECOND examples, the SVM achieves 94.3% accuracy on the test set.

Manshadi and Allen (2011) build on top of previous work by proposing an approach for "unrestricted" quantifier scope disambiguation. Specifically, they go beyond the case of two quantifiers per sentence with a method that can deal with an arbitrary number of quantifiers per sentence. They do this by addressing each individual pair of quantifiers, determining if one outscopes the other, and if so, which one outscopes the other. With this, they represent the quantifier scope disambiguation of a sentence as a directed acyclic graph (DAG) and use a graph similarly metric to measure the closeness of a predicted graph to the gold graph. Their dataset consists of 500 sentences manually extracted from the web.

The issue with the above approaches is that QSD is not a problem that can be resolved simply by considering surface level patterns. Rather, commonsense reasoning and world knowledge is often required to disambiguate the sentences in question. So then, the issue becomes how to design an approach that takes this world knowledge into account and uses it to reason about the possible interpretations of an ambiguous sentence.

Saba and Corriveau (2001) provide a framework for the incorporation of commonsense reasoning into the resolution of quantifier scope ambiguities. They hold that disambiguation requires reasoning about the expected size of each quantified entity when it enters into a relation with another quantified entity. They formalize this with a notion they call a "quantificational constraint", which takes two concepts C_1, C_2, as well as a relation R, and outputs numbers m_1, m_2 which relate elements in C_1 to C_2. Specifically, m_1 is the number of C_1 elements that are in relation R to the same C_2 element, and m_2 is the number of C_2 elements that are in relation R to the same C_1 element. They provide the following example: given the two concepts $C_1 =$ "house" and $C_2 =$ "street" and the relation $R =$ "on", we have $m_1 = \textit{many}$ and $m_2 = 1$. This is because we can have many houses on the same street, but a house cannot be located on more than one street. With quantificational constraints, they devise a method to measure the plausibility of each possible reading of a sentence with ambiguous quantifier scope. However, they do not perform any empirical validation of their method.

Srinivasan and Yates (2009) seek to explicitly inject pragmatic knowledge into a QSD model. They use corpus statistics to estimate sizes of sets that enter into relations with each other. For example, we understand that a doctor tends to live in one city, rather than living in multiple cities simultaneously. And thus, in the sentence "There is a doctor for every city", we would take the quantifier "every" to take wide scope. The authors integrate this pragmatic knowledge into a Markov Logic Network (MLN). They have very small training and test sets for the MLN, and they make use of the Web1Tgram corpus from Google to obtain corpus statistics.

AnderBois et al. (2012) approach the problem of QSD from a linguistics perspective, seeking to identify the most important factors leading to a resolution decision. They find that the linear order of quantifiers (which comes first, second, etc.), the grammatical function of the constituents the quantifiers are located in (subject or object), and the lexical realization (identity) of the quantifiers are all important factors. They do not consider world knowledge and commonsense reasoning in this work, but acknowledge that it must also be an important factor. They work with a dataset of 358 clauses containing two quantifiers from LSAT logic games questions. They choose to work with logic games data to avoid dealing with examples that require commonsense reasoning to resolve.

We were able to get in touch with one of the authors of the above work, who referred us to

Justyna Grudzińska, Aleksander Wawer, and Marek Zawadowski. They are presently working on integrating new predictors of quantifier scope into a machine learning model. They add to the work of AnderBois et al. (2012) by providing insights into the effect of prepositions on quantifier scope. Certain prepositions, such as "with" tend to block the inverse scope reading, while other prepositions, such as "of" and "to", enable it (though they do not require it). Their experiments are also conducted on the small LSAT logic games corpus. They find success by concatenating universal sentence encoder representations (Cer et al., 2018) with their features and feeding this into a support vector machine. Unfortunately, experiments with BERT were less successful. This is likely because the LSAT logic games dataset is far too small to be used as a fine-tuning set. Justyna, Aleksander, and Marek agree that QSD is a very hard problem and that more data is required to see greater success with these machine learning models.

Work has also been done on understanding how humans process quantifier scope ambiguity. Feiman et al. (2020) replicate an earlier study by Chemla and Bott (2015) on the priming of quantifier scope preferences and conduct new experiments. They are unable to find definitive evidence for or against the claim that the inverse scope reading can be primed. Similarly, there is no evidence for or against the priming of a "U-wide" reading (the reading where the universal quantifier takes wide scope).

Dwivedi (2013) focuses on the levels of processing that humans use to interpret sentences with ambiguous quantifier scope. They find that these sentences tend to be processed at a heuristic level, which is informed by lexical-pragmatic biases humans possess based on the NPs and verb being used. People only exhibit a deeper ("algorithmic") level of processing when asked specifically about which reading is the correct one. Moreover, people tend to exhibit a very strong preference for the surface scope reading, which is in line with labelled corpora for quantifier scope.

2 Examples

We now take the time to break down specific examples of quantifier scope disambiguation to get a better feel for the problem we aim to solve.

Consider the following example from the dataset of Andrew and MacCartney (2004):

1. Each chair is occupied by exactly one of the diplomats.

There are two possible readings depending on which quantifier is interpreted to have wide scope.

FIRST reading: $\forall$ chair $\exists!$ diplomat: occupy(diplomat, chair).

SECOND reading: $\exists!$ diplomat $\forall$ chair: occupy(diplomat, chair).

The two readings do not necessarily have to contradict each other. There can be a single diplomat d such that d is the sole occupant of every chair. In this case, both readings are true. However, this is not a very plausible situation. Furthermore, under the second reading, it is impossible that each chair is occupied by a different diplomat. This is problematic, since the latter is by far the most plausible scenario. Thus, we should prefer the first reading.

How can we design a system which can learn to reason in the way that was exhibited above? It's clear that the reasoning is probabilistic in nature, as we reflect on which of the two scenarios is more likely, given our world knowledge. For example, we know that it is common for an individual to sit on a chair, but it is not common for a person to sit on multiple chairs at the same time. This links back to the ideas in Saba and Corriveau (2001) and Srinivasan and Yates (2009) of estimating sizes of sets that enter into a relation. Here, we are considering the set of people, the set of chairs, and the relation of "occupying", which we can infer to be synonymous with "sitting" in this case. So, the mental procedure that we follow to disambiguate this sentence is to first map out the possible readings, then pick the most sensible one based on our world knowledge about set sizes.

Let us try this with another example.

2. Exactly one of the officials must be assigned to each court.

FIRST reading: $\exists!$ official $\forall$ court: assignedTo(official, court).

SECOND reading: $\forall$ court $\exists!$ official: assignedTo(official, court).

This example is isomorphic to the first example. We have essentially the same situation, but here the quantifiers appear in a different order. Thus, we prefer the second reading here.

Once again, we reason about each of the possible readings that we have extracted, selecting the most plausible one. In tennis tournaments, each court is assigned a (different) official. We understand that it is unrealistic for one official to simultaneously be responsible for multiple courts. Thus, the second reading should be assigned a higher probability than the first reading.

The dataset also contains some peculiar examples involving negation.

3. No one who leaves the elevator returns to it on a different floor.

FIRST reading: $\nexists$ person: $\exists$ floor: leaveElevator(person) $\wedge$ ReturnToElevatorOn(person,floor)

$\iff$ $\forall$ person: $\neg$ [$\exists$ floor: leaveElevator(person) $\wedge$ ReturnToElevatorOn(person,floor)]

$\iff$ $\forall$ person $\forall$ floor: $\neg$[leaveElevator(person) $\wedge$ ReturnToElevatorOn(person,floor)]

SECOND reading: $\exists$ floor: $\nexists$ person: leaveElevator(person) $\wedge$ ReturnToElevatorOn(person, floor)

$\iff$ $\exists$ floor $\forall$ person: $\neg$[leaveElevator(person) $\wedge$ ReturnToElevatorOn(person,floor)]

In this case, the second reading does not make much sense because of the presence of the relative pronoun "who". Since the quantifier "a" is situated in the relative clause starting with "who", it becomes clear that "No" should be the quantifier taking wide scope.

This is a particularly interesting example because reasoning based on set sizes does not work here. If anything, the reasoning here is based on syntax rather than on pragmatics, with the relative clause being the biggest indicator that the first quantifier is the one that takes wide scope. Furthermore, both readings are totally plausible here.

The following is an example involving two universal quantifiers:

4. Each of them enters all three events.

FIRST reading: $\forall$ person $\forall$ event: enter(person, event).

SECOND reading: $\forall$ event $\forall$ person: enter(person, event).

Here, both readings are equivalent. However, strangely, the gold label for this example is that the FIRST reading is preferred. Am I missing something here?

The following is an example involving two existential quantifiers:

5. At an evening concert, exactly six songs will be performed.

FIRST reading: $\exists$ concert $\exists$ $s_1, \ldots, s_6$: performedAt($\{s_1, \ldots, s_6\}$, concert).

SECOND reading: $\exists$ $s_1, \ldots, s_6$ $\exists$ concert: performedAt($\{s_1, \ldots, s_6\}$, concert).

Once again, the two readings are equivalent. And yet, the gold label for this example is that the FIRST reading is preferred. Once again, am I missing something here?

6. No two of these live in a single house.

FIRST reading: $\nexists$ $x_1, x_2 \in$ “these”: $\exists$ house: liveIn($\{x_1, x_2\}$,house).

$\iff$ $\forall$ $x_1, x_2 \in$ “these” $\forall$ house: liveIn($\{x_1, x_2\}$, house).

$\iff$ $\nexists$ house: $\exists$ $x_1, x_2 \in$ “these”: liveIn($\{x_1, x_2\}$, house).

SECOND reading: $\exists$ house $\nexists$ $x_1, x_2 \in$ “these”: liveIn($\{x_1, x_2\}$, house)

This is equivalent to Example 3, and once again the first reading is preferred. Interestingly, this is another scenario where both readings are plausible. Here, the main driver behind this reading is the syntax of the sentence. Ironically, the heuristic of selecting the quantifier that appears first is a driving force is the disambiguation decision here. Another way to reason about this is that if the existential quantifier were to take wide scope, then the word “the” would have been used as a way of referring to a specific house that does not contain two of these.

7. Directly connected subway stops are those stops between which there is a subway connection that passes through no other stop on the way from one to the other.

FIRST reading: $\exists$ connection $\nexists$ other stop: passesThrough(connection, other stop).

SECOND reading: $\nexists$ other stop $\exists$ connection: passesThrough(connection, other stop).

$\iff$ $\forall$ other stop $\neg$ [$\exists$ connection: passesThrough(connection, other stop)].

$\iff$ $\forall$ other stop $\forall$ connection: $\neg$ [passesThrough(connection, other stop)].

Here, the first reading is preferred. The second reading requires that every connection between the two stops does not pass through any other stop, which is too strong of a statement. Conveniently, the wording of the sentence provides a major clue as to what the preferred reading is. The use of the term “there is” (which translates to “there exists”) right before the first quantifier “a” tells us that the existential quantifier should take wide scope.

8. For each of the three areas, none of the five employees receives the same ranking.

FIRST reading: $\forall$ area $\nexists$ e_1, e_2: rankingIn(e_1, area) $=$ rankingIn(e_2, area).

$\iff$ $\forall$ area $\forall$ e_1, e_2: rankingIn(e_1, area) $\neq$ rankingIn(e_2, area).

SECOND reading: $\nexists\ e_1, e_2\ \forall$ area: rankingIn(e_1, area) = rankingIn(e_2, area).

$$\iff \forall\ e_1, e_2\ \neg\ [\forall \text{ area: rankingIn}(e_1, \text{area}) = \text{rankingIn}(e_2, \text{area})].$$

$$\iff \forall\ e_1, e_2\ \exists \text{ area: rankingIn}(e_1, \text{area}) \neq \text{rankingIn}(e_2, \text{area}).$$

In this example, the first reading is preferred. As we have seen in previous examples, this is a case where the first reading implies the second reading, but the latter is a weaker statement. Like the previous example, the natural language statement sounds similar to an FOL statement. Beginning the sentence with "For each" (which is equivalent to "for all") is a strong indicator that the universal quantifier takes wide scope here.

3 Existing Datasets

We obtained the dataset from Higgins and Sadock (2003) and filtered out examples where there is no scope interaction between the two quantifiers. We also filtered out examples containing two existential or two universal quantifiers. We were left with only 21.1% of the data. On our filtered set, predicting the first quantifier works 72.3% of the time (we did not split into training and test sets). Predicting the universal quantifier only works 59.0% of the time. In the other 41.0% of cases, we often see the quantifiers "some", "many", "most", and "another" taking wide scope over the universal quantifiers. Interestingly, in the case where the second quantifier takes wide scope, it is a universal quantifier in 88.5% of cases. By contrast, when the first quantifier takes wide scope, it is universal in 47.8% of cases. These findings inspired us to try an additional experiment. In this case, we predict that the first quantifier takes wide scope if it is a universal quantifier OR if it is "some", "many", "most", or "another". Otherwise, predict that the second quantifier takes wide scope. This results in 79.8% accuracy. (Of course, since there is no test set, this is "cheating" somewhat.)

We also currently have access to the dataset of Andrew and MacCartney (2004). Unfortunately, it is small in size, containing less than 300 examples. Furthermore, in many examples, both quantifiers are of the sample type (universal or existential), and yet the gold label stats that one outscopes the other. This should not be possible, since two quantifiers of the same type are interchangeable. For the purposes of our study, we filter out these examples, focusing only on sentences with one universal and one existential quantifier. Moreover, we eliminate sentences where neither quantifier outscopes the other (for now). This leaves us with only half of the original dataset. More specifically, after filtering, we have 133 training examples and 21 test examples. We find that simply predicting that the first quantifier takes wide scope has 88.7% accuracy on the training set, and 71.4% accuracy on the test set. Meanwhile, predicting that the universal quantifier takes wide scope has 89.5% accuracy on the training set, and 100% accuracy on the test set. This reflects the need for a much larger dataset for quantifier scope disambiguation, where these heuristics cannot be relied upon to obtain good performance. This will not be easy, however, as it has not been attempted before and it is currently unclear as to how we can generate examples for this problem from a template.

We start by controlling for linear order and quantifier type (universal/existential) in the datasets of Higgins and Sadock (2003) and Andrew and MacCartney (2004). That is, we require the following:

$$\#\ (\text{first, universal}) = \#\ (\text{first, existential}) = \#\ (\text{second, universal}) = \#\ (\text{second, existential}) = n \tag{1}$$

This implies that a model that always prefers the first quantifier to take wide scope would achieve chance performance. The same would happen for a model that always prefers the universal quantifier. It is our hope that this would either encourage models to perform pragmatic inference or, more likely, expose models' inability to perform pragmatic inference.

However, the filtered training set from Higgins and Sadock (2003) is unbalanced, especially when it comes to the case where the second quantifier takes wide scope:

	First	Second
Universal	65	46
Existential	71	6

We list the six sentences where the second quantifier is existential and takes wide scope:

- Behind all the hoopla is some heavy-duty competition.
- Investors in all funds will seek safety in the coming months, some analysts say.
- There has been no announcement of the Burger King arrangement by either party, possibly for fear that McDonald's and other fast-food rivals would seize on it in scornful advertising.
- Ordinarily in genetic engineering each of these genes, minus the one that caused the virulence, would have been transferred to another bacterium, called E. coli, which would then produce a nonvirulent version of the toxin.
- All the tie-ins, though, have some marketing experts questioning whether the party may go too far.
- They then had no choice in many cases but to sell the contracts at prevailing prices – in most cases at a substantial loss.

Here are some examples of sentences where the first quantifier is universal and takes wide scope:

- The materials in each set reach about 90 students.
- A telephone-information operator had no listing for either party.
- Where truly representative governments are safeguarded by constitutional protections of human rights and an independent judiciary to construe those rights, there is no excuse for breaking the law because some individual or group disagrees with it.
- If there are any signs that Mr. Major will be less inclined to use interest-rate boosts to rescue the pound from another plunge, that currency is expected to fall sharply.
- The state could also increase gasoline taxes; every one penny increase in the tax would yield $11 million a month.

Here are some examples of sentences where the first quantifier is existential and takes wide scope:

- What makes the most sense is to find someone who wants to buy the whole company or cause a recapitalization of all shares.
- Some analysts don't expect a quick revival of any takeover by the pilots.

- According to some analysts familiar with the negotiations, the 10% of equity would come directly from Donaldson Lufkin and a fund affiliated with the investment bank Blackstone Group, which would reduce their CNW equity holdings by 5% each.
- So it will take many quarters for IBM to roll out all the products that customers need, and it will take years for customers to integrate the products into their operations.
- Some companies were delinquent in filings and other actions, all of which cost money, Mr. Holmes said.

Many of the sentences where an existential quantifier takes wide scope contain phrases like “some say” or “someone who”.

Here are some examples of sentences where the second quantifier is universal and takes wide scope:

- Minpeco now says it is willing to settle for up to $65.7 million from each brother, although the actual amount would probably be much less.
- A quarter of a million people cross the Bay Bridge every day, far more than the 100,000 that use the Bay Area Rapid Transit system – which was working but wasn’t stopping in the city’s Financial District yesterday afternoon because electricity was shut off and the area was being checked for gas leaks.
- And though federal law dictates that only $100 million can be disbursed from that fund in any one state per disaster, administration officials expect Congress to move in to authorize spending more now in California.
- Now there are many cars for every purse and purpose.
- After only 11 moves for each side, the computer’s position was shaky.

We sample from our filtered dataset to produce a balanced dataset which satisfies Equation (1). Since # (second, existential) = 6 for the filtered set, we have $n = 6$. As a consequence, our balanced dataset consists of 24 examples. This is very small, but can potentially serve as a starting template for the generation of new examples.

The filtered training set from Andrew and MacCartney (2004) is even more unbalanced than Higgins and Sadock (2003). There is only one example where the second quantifier is existential and takes wide scope. It is impractical to make a balanced dataset from this, and so we do not do so. Furthermore, we are unclear on whether or not the authors had a license to use logic puzzles examples, and their paper was never published.

	First	Second
Universal	105	14
Existential	13	1

4 Alternative Problem Formulations

The simplest formulation of the QSD problem is as a binary classification task, where, given a text span containing two quantifiers, the goal is to correctly determine whether the first or second takes wide scope. Additional labels can also be included, such as for the case where neither quantifier outscopes the other. However, given the lack of a large standard dataset for this problem, it may be preferable to fit QSD into the framework of an existing popular NLP task. If the latter is possible, we can easily evaluate many existing NLP models on QSD.

4.1 Natural Language Inference

In an attempt to fit our problem of quantifier scope disambiguation into the context of large pre-trained language models, which have come to dominate the field of NLP, we formulate QSD as a natural language inference task. We do this by casting a sentence with ambiguous quantifier scope as the "premise" sentence and generating multiple "hypothesis" sentences, where each one designates a different reading of the ambiguous premise.

The benefit of this approach is that large pre-trained language models such as BERT can easily be evaluated, assuming that an NLI-like dataset for QSD can be generated (we will shortly see that this is not trivial). One can simply obtain a pre-trained BERT easily[1] and can fine-tune it on MNLI (Williams et al., 2018). Indeed, Jeretic et al. (2020) use this approach in evaluating BERT on scalar implicature and presupposition.

In our particular case, we are not really concerned with distinguishing between the "contradiction" and "neutral" labels. Rather, we wish to know whether or not a particular reading is implied by the ambiguous sentence.

We illustrate the NLI formulation with the first example from Section 2:

(a) *Each* chair is occupied by *exactly one* of the diplomats. Each chair is occupied by the same diplomat. (Label: Not implied)

Perhaps this is not the best formulation. What if we try to stay closer to the FOL interpretations of the sentence?

(b) *Each* chair is occupied by *exactly one* of the diplomats. There is one diplomat who occupies every chair. (Label: Not implied)

But is the second sentence in the above truly unambiguous? We argue that it is because of the presence of the relative pronoun "who".

It is clear that the difficulty with designing this task is being able to come up with an unambiguous "hypothesis" sentence that is clearly aligned with one of the two readings that we express in FOL. We do not want to have a hypothesis sentence that is simply another multi-quantifier sentence that paraphrases the first sentence.

So far, we have written two hypotheses that are in line with the second reading. What about a hypothesis that is in line with the first reading?

(c) *Each* chair is occupied by *exactly one* of the diplomats. For each chair, there is one diplomat who occupies it. (Label: Implied)

From this example alone, we can see that generating hypothesis sentences is highly non-trivial. It is unlikely that this can be done automatically. This represents a major roadblock to this approach. Nonetheless, we manually construct hypothesis sentences for the examples in Section 2. We then inspect the output of BERT-base fine-tuned on MNLI. We find that BERT always outputs "Implied"

[1] https://huggingface.co/transformers/model_doc/bert.html

(in MNLI terms, “Entailment”), and thus completely fails on this task. It is very likely that BERT’s failure is due to its heavy reliance on the lexical overlap heuristic (McCoy et al., 2019).

4.2 Question Answering

Question answering is another extensively studied NLP task. We briefly consider the possibility of formulating QSD as a question answering problem. This can be done by asking questions regarding the interpretation of the ambiguous sentence. This is best illustrated by an example. For the sentence “Each chair is occupied by exactly one diplomat”, we might ask the following questions:

- Does the same diplomat occupy every chair?
- Is there exactly one diplomat present?

However, we can already see that this formulation is isomorphic to the NLI formulation. We can map any question we generate to a hypothesis sentence in a one-to-one fashion. Hence, the same issues that plague the NLI formulation are also present here.

In addition, the above formulates QSD specifically as a yes/no question answering problem. This does not fall into the category of extractive question answering, where the answer can be directly extracted word-for-word from the text passage. Because of this, it is not possible to make use of BERT fine-tuned on SQuAD (Rajpurkar et al., 2016). Instead, we could use BERT fine-tuned on the recent BoolQ dataset for yes/no question answering (Clark et al., 2019). However, we remain likely to fall into the same trap as in the case of MNLI. Therefore, we abandon this course of action and we concentrate our efforts on the construction of a large dataset for QSD that falls under the original problem formulation.

5 Dataset Construction

It is clear that a much larger amount of data than what is currently available is required to have a chance at solving this problem. Hence, in this section, we assess possible methods for constructing a new dataset for quantifier scope disambiguation.

5.1 Generation from Templates

One automatic method to construct a dataset for this problem would be to generate examples from pre-specified templates, which specify a sequence of constituents.

For example, drawing inspiration from the “Each chair is occupied by exactly one diplomat” example, we try to develop a general formulation that can be used to generate more examples like it. We may have a template like the following:

$$\underline{\text{Each}}\ [\text{NP}_1]\ [\text{AUX}]\ [\text{V}]\ [\text{P}]\ \underline{\text{exactly one}}\ [\text{NP}_2]. \tag{2}$$

From this, we can generate examples like:

- Each chair is occupied by exactly one diplomat. (FIRST)
- Each chair is destroyed by exactly one diplomat. (SECOND)
- Each chair is placed at exactly one table. (FIRST)

Or, perhaps we can try something simpler like

$$\underline{\text{Each}}\ [\text{NP}_1]\ [\text{V}]\ \underline{\text{a}}\ [\text{NP}_2]. \tag{3}$$

We can generate the following:

- Each participant signs a waiver. (SECOND)
- Each participant brings a waiver. (FIRST)

It may actually be possibly for us to generate for templates in this way. The problem is that there is no clear-cut rule for how to label each example. There is no template for commonsense reasoning. At a minimum, the label depends on the first quantifier, the first noun, the verb, the second quantifier, and the second noun. We cannot hope to write rules for every possible (Q_1, N_1, V, Q_2, N_2) tuple. The only option seems to be to crowdsource the annotation of the generated examples. But what will inter-annotator agreement look like? Previous attempts at constructing a dataset for QSD have reported low inter-annotator agreement (Higgins and Sadock, 2003; Manshadi and Allen, 2011).

5.2 Scraping Existing Corpora

A second approach is to scrape existing corpora for relevant examples. This has the benefit that we would be extracting naturally occurring sentences with quantifier scope ambiguity.

Note: It is not immediately clear how to extract quantifiers from a parse. We can parse corpora (or use pre-parsed corpora like PTB WSJ) and extract sentences that contain two quantifier phrases (constituents labelled as QP). However, not all quantifiers are encoded as "quantifier phrases", so the proposal above is unlikely to work. For instance, the quantifier "no" is designated as a determiner. A potential solution to this problem is to forgo parsing altogether and instead have a list of quantifiers. We search for sentences which contain two of the quantifiers. We can further specify which quantifiers are universal and which are existential. Thus, we ensure that each sentence we extract contains one universal and one existential quantifier.

Unfortunately, there are several drawbacks to consider with the scraping approach. First, many of the two-quantifier sentences are likely to have no scope interaction. It is frequently the case with these sentences that each quantifier is located in its own separate clause. Recall that 61% of the sentences in the Higgins and Sadock (2003) corpus, which comes from PTB, have no scope interaction between the two quantifiers. Second, a related problem is that we are very likely to extract far more sentences where the first quantifier takes wide scope (surface scope interpretation) than where the second one does (inverse scope interpretation). Indeed, this disparity manifests itself in all existing corpora. AnderBois et al. (2012) remark that linear order of quantifiers is a key factor in quantifier scope disambiguation, with the first quantifier taking wide scope being the most common resolution. Third, quantifier scope ambiguity is a rare phenomenon, and thus it may be difficult to extract a large dataset from a text corpus, especially given that we are restricting to the case of sentences with two quantifiers that exhibit scopal interaction.

As a preliminary experiment and as an attempt to assuage the concerns above, we extract relevant sentences from a portion (about 1/30) of the Gigaword 3 (Graff et al., 2007) + Wikipedia 2018 corpus that was used to train the Hilbert-MLE embeddings (Newell et al., 2019). We naively select all sentences containing one universal and one existential quantifier. Our list of universal quantifiers is: all, any, both, each, no, none, neither, every. Our list of existential quantifiers is:

exactly, some, a, an, either, many, most. (For simplicity, we start with one-word quantifiers). Out of 7.76 million sentences, we extracted 448k sentences with our simple filtering method. Here is a sample of extracted sentences (along with a hand-annotated label we provide after choosing the example to put in the list):

- When a playable character loses all hit points, he or she faints. (LABEL: FIRST)
- The Baire Category Theorem says that every complete metric space is a Baire space. (Label: FIRST)
- Each phase has a characteristic arrangement of atoms. (Label: FIRST)
- Every real number has a unique location on the line. (Label: FIRST)
- There, it is called ketos, a term that initially included all large marine animals. (Label: FIRST)
- Most toothed whales have no fixed bonds. (Label: FIRST)
- Governments can play a role in all of these areas. (Label: SECOND)
- In magnetostatistics, such vector fields model static magnetic fields on a region of the plane containing no current. (Label: FIRST)
- However, the battle was in vain, as neither platform captured a significant share of the world computer market and only the Apple Macintosh would survive the industry-wide shift to Microsoft Windows running on PC clones. (Label: FIRST)

This sample suggests that FIRST is the most likely label, which would be in line with the literature. Because of the difficulty of obtaining gold annotations, we cannot speak for the entire set.

Many of our sentences have no scoping relation between the two quantifiers. This is expected, and we now have to think about a way of detecting these automatically, if possible. Far too many examples have intervening punctuation marks, quotations and/or conjunctions between the two quantifiers. Indeed, while manually searching the extracted sentences, it is difficult to find a sentence which does not have this problem. Thus, in a second attempt, we impose an additional filter to remove these cases. This is in line with what was done by AnderBois et al. (2012). From 448k sentences, this filtering step brings us down to 216k sentences. Here are some examples:

- Because all living beings possess a soul, great care and awareness is essential in one's actions. (Label: FIRST)
- Each night a jury was selected from members of the audience; based on the jury's vote, one of two different endings would be performed. (Label: FIRST)
- All this can happen in about a day. (Label: SECOND)
- These brought with them many new words to the European vocabulary for which there was no previous Latin equivalent. (Label: FIRST)
- Summers are moderately warm with a number of hot days every month.

However, there are still many examples that would not be suitable for a QSD dataset:

- Alabama, along with Oklahoma, has the most reported EF5 tornadoes of any state, according to statistics from the National Climatic Data Center for the period January 1, 1950, to June

2013.

- It is focusing both on a single person and the whole community.
- In the final chapters of the novel, he suffers a complete mental breakdown upon realizing that he can no longer deceive himself in this respect.
- Commonly the effect of animation is achieved by a rapid succession of sequential images that minimally differ from each other.
- Agassi, along with five athlete partners (including Wayne Gretzky, Joe Montana, Shaquille O'Neal, Ken Griffey Jr., and Monica Seles) opened a chain of sports-themed restaurants named Official All Star Cafe in April 1996.
- It is considered one of the most significant rally cars of all time, because it was one of the first to take advantage of the then-recently chagned rules which allowed the use of four-wheel drive in competition racing.
- Most scholars reject these accounts as no more than hearsay, preferring instead the account given by Attila's contemporary priscus.

From these examples, we observe that there are various ways in which there can be no scoping relation between two quantifiers.

In some cases, only part of a quantifier is extracted because of how we performed the extraction. For example, consider the sentence (which certainly does not apply to Montreal):

- Summers are moderately warm with a number of hot days every month.

Our automatic extractor pinpoints "a" and "every" as the quantifiers in this sentence. However, the first quantifier should really be "a number of".

To resolve some of the above issues, some of the filters we can try include:

- Narrow it down to the case where one quantifier is in the subject and the other is in the object. (One naive filter could be to parse the sentence and ensure that there is a verb between the two quantifiers. However, this would eliminate some of the valid sentences that we have.)
- Make use of constituency parse features (unclear how at the moment).
- Remove examples containing the word "most" when it is immediately preceded by the word "the".
- Remove examples containing the word "each" when it is immediately followed by the word "other".
- A yet-to-be-determined filter to deal with the case of "both". (Perhaps based on its POS tag?)
- Filter out intervening parentheses ( ).
- Filter out cases with the terms "all but", "nearly all", "almost all", etc.

Applying the filter for "most" and "each" brings us from 216k examples to 211k examples.

To achieve gold annotations for our extracted sentences, we will likely have no choice but to do crowdsourcing. This can be done via a platform like Amazon Mechanical Turk or by asking for volunteers in-house at Mila (not sure if we have enough volunteers, however). A reasonable target would be for at least three annotators to see each example. One possible annotation scheme could be to use a numeric scale from 1 to 5, where

1) It is clear that the first quantifier takes wide scope.
2) The first quantifier is likely to take wide scope.
3) The sentence is too ambiguous to resolve.
4) The second quantifier is likely to take wide scope.
5) It is clear that the second quantifier takes wide scope.

Alternatively, we can simply give three options: the first quantifier takes wide scope, the second quantifier takes wide scope, or the ambiguity is not resolvable / there is no scopal interaction. Regardless of which annotation scheme we choose, we may also want to give annotators the option of changing what in a sentence is designated as a quantifier.

Before we move to annotations, we must first decide on an automatically extracted set of examples that we want to put forward to annotators. Even after our naive attempt, there is still much to consider. We only worked with 1/30 of the corpus and still managed to extract over 200k sentences (pending more aggressive filtering). Without human annotations, it is unclear how many of these sentences would be valid examples, and we can only rely on samples we draw from our set of sentences. With all of this mind, here are some next steps and outstanding items to consider:

- Do we need to extract from additional corpora?
- Should we avoid English Wikipedia since this is often used to train large LMs?
- How do we incorporate additional quantifiers beyond our current (relatively) small list?
- How can we ensure and verify that the vast majority of our extracted sentences are valid (that is, they contain ambiguous quantifier scope) and resolvable? (We will likely have to draw a sample of size 50-100, manually annotate, and then extrapolate).
- How can we control the distribution of surface scope readings vs. inverse scope readings? What about universal wide vs. existential wide? (Or perhaps we don't control for this and have first-second and universal-existential F1 scores?)

5.2.1 Annotating a Sample

Before proceeding any further, we take a few steps back to better understand the data that we are extracted. This will help us better decide which filters to use and when to proceed to human annotation.

We work with the same 1/30th of the Gigaword 3 + Wikipedia 2018 corpus. We modify our list of quantifiers slightly as well. For this next part, our universal quantifiers are: *all*, *any*, *both*, *each*, *every*, *everyone*. Our existential quantifiers are: *some*, *a*, *either*, *an*, *someone*. We removed the negative universal quantifers *no*, *none*, and *neither*, but we intend on re-introducing them later. We also removed *most* and *many* from our list of existential quantifiers, as it is not entirely clear if they can be treated as existential quantifiers. Unlike the quantifiers *some* and *a*, *most* and *many* depend on the size of the set that they are being applied to.

With our modified list of quantifiers, we apply filtering and obtain 93.7k sentences. We leave the filter for intervening colons, semicolons, quotes, and conjunctions on, but we do not deal with other intervening punctuation for the time being. It was decided that filtering out cases with intervening commas would be too harsh, and similarly for periods (consider the case of "Dr."). We do however continue to apply the filter for "each other".

From the 93.7k sentences, we sample 100 sentences at random and manually label them. We use four labels:

- FIRST: The first quantifier takes wide scope.
- SECOND: The second quantifier takes wide scope.
- NO: Neither quantifier outscopes the other.
- INVALID: The quantifiers were not correctly identified by the automatic extractor.

The following is a list of observations made during the manual annotation process:

- Some examples are missing tokens (usually they are numeric tokens). This probably resulted from an issue in the tokenization process. An example of this is the instance "Trains average with a maximum of on all but the automated driverless trains of line 14, which average and reach ".Regardless, it was always possible to interpret the example.
- There is a sentence that repeats itself (with some slight variation) multiple sample. It has the form "According to the United States Census Bureau, the city has a total area of x, all of it land". A similar repated sentence has the form "According to the 2010 census, x has a total area of y, all land".
- There are many sentences where there are more quantifiers than the two that were identified by the extractor. This is because our list of quantifiers is very small. We do not capture "most", "a number of", "one of", "half a dozen", "three", "12", "over 130", etc. There are also cases where there is a multi-word quantifier and the extractor only identifies the word "a" (e.g. the extractor identifies "a" as a quantifier, when really it should be "a number of").
- Events that occur annually appear several times in the sample. That is, examples like "Each year, a festival is held". How do we interpret "a festival"? Is each instance of the festival treated separately or do we view it simply as the same annual festival?
- There are examples where the resolution is much more obvious, while others require a lot more thinking (we needed to write down the sentence in FOL to be sure and even then, it's still tough). (e.g. "Current biology and ecology textbooks use the latter "De Bary" definition, or <u>an</u> even broader one where symbiosis means <u>all</u> interspecific interactions; the restrictive definition where symbiosis means only mutualism is no longer used.")
- We did not account for intervening "–", which leads to a label of NO in all cases.

In short, manual annotation for this problem is highly non-trivial. The issues mentioned above illustrate that well. We were unsure of how to annotate many of the examples. We suspect that inter-annotator agreement for a dataset like this would be very low.

After removing repeated sentences, we are left with 88 sentences. We obtain the following results from our annotation:

Label	**Number of Examples**
First	43
Second	14
No	22
Invalid	9

Encouragingly, we find that in 65% of the examples, one of the quantifiers outscopes the other (at least, based on how we annotated). It should be noted however that for many examples, it was not immediately obvious whether we should label as First/Second or as No. In line with existing QSD datasets, we find an 76:24 ratio between the First and Second label. Thus, we retain the strong class imbalance while also maintaining representation from the minority class.

	First	Second
Universal	19	2
Existential	24	12

Unlike the other QSD datasets, the existential quantifier takes wide scope more often than the universal quantifier does. This is quite a surprising result. It should also be noted that in 35 out of the 36 times an existential quantifier takes wide scope, that quantifier is the indefinite article "a(n)".

Another one of our concerns was whether or not to filter out cases with a comma in between the two quantifiers. Our results suggest that we should not apply such a filter:

Label	**Number of Intervening Commas**
First	9 (20.5% of all "First")
Second	2 (14.3% of all "Second")
No	11 (47.8% of all "No")

5.3 Data Augmentation

The results obtained from our sample are very promising. However, it is still expensive to obtain human annotations for hundreds to thousands of examples. With this issue in mind, we explore data augmentation strategies that may allow us to construct a large dataset from a small set of human-annotated examples. We also seek to augment the data in such a way that we can generate more examples for the minority class (SECOND).

5.3.1 Quantifier Switching

We begin by exploring very simple methods. In a naive attempt, we simply switch the two quantifiers in each sentence from our sample. In our annotation of the new "flipped" examples, we find that the vast majority of them are nonsensical. We did not even finish annotation as the pattern was clear after annotating nearly 20 examples.

Below is one of the many examples where switching quantifiers fails to produce a sensible sentence:

- Original sentence: The commissioners are elected county-wide, in staggered terms, and each serves a four-year term.

- “Flipped” sentence: The commissioners are elected county-wide, in staggered terms, and a serves each four-year term.

5.3.2 Sentence Passivization

In our next attempt, we seek to passivize sentences in active voice to see if this changes the interpretation. We hypothesize that such an approach can be effective and will lead to an inversion of the label (FIRST to SECOND, SECOND to first) if one quantifier is in the subject and the other is in the object. However, many instances in our sample do not take this form. In this case, passivization results in the quantifiers retaining the same linear order as before, or passivization is not possible to begin with. Nonetheless, our hypothesis was validated, as we do find that in cases where one quantifier is in the subject of a verb and the other is in the object of the same transitive verb, passivization leads to an inversion. One thing to keep in mind going forward is that it is unclear how to automatically passivize (perhaps this can be done with the help of a parse). If this can be achieved, then this could be a straightforward data augmentation strategy. The warning would just be that we can’t apply this to every instance in a seed set, but rather only to those of a specific form. (Note: We can also turn passive voice sentences into active voice to obtain an inversion. We observe that there are quite a few passive voice sentences in our sample.)

Below is an example of an instance where passivization is possible and leads to a label change:

- Original sentence: The commissioners are elected county-wide, in staggered terms, and each serves a four-year term. (Label: FIRST)
- Passivized sentence: The commissioners are elected county-wide, in staggered terms, and a four-year term is served by each. (Label: SECOND)

Here is a case where passivization does not lead to a label change (since the linear order of the quantifiers does not change):

- Original sentence: Shamans (curanderos) played a pivotal role in social relations in both groups. (Label: FIRST)
- Passivized sentence: A pivotal role in social relations in both groups was played by shamans (curanderos). (Label: FIRST)

And finally, we show an example of a sentence that cannot be passivized:

- Both titles of an early story by John Crowley, first published in 1978 as “Where Spirits Gat Them Home”, later collected in 1993 as “Her Bounty to the Dead”, come from “Sunday Morning”.

5.3.3 Synonym Replacement

Wei and Zou (2019) explore four simple data augmentation operations for text classification datasets. Though their operations will not help us change the label distribution, they are worth considering. These methods are: synonym replacement (replace n words with their synonyms), random deletion (randomly delete words with probability p), random swap (randomly swap the order of two words in the sentence), random insertion (randomly insert synonyms of a word at a random position). We remain sceptical that the later three operations will be effective, as there is no guarantee that the resulting sentences will be coherent. We are more optimistic that synonym replacement can be effective.

We apply the synonym replacement operation to the sentences in our sample by replacing 20% of the words in each sentence with one of their synonyms, as given by WordNet (Miller, 1998). We make sure of an implementation found online[2]. Unfortunately, this data augmentation strategy was not as successful as we hoped.

- Original sentence: This started a tradition since continued by every subsequent Stanley Cup champion.
- Modified sentence: This started a tradition since go along by every subsequent Stanley Cup admirer.

- Original sentence: In 1563, Norwegian troops stopped the Swedish advance at Elverum, which provided a strategic point since it lay on both north-south and east-west trade and travel routes.
- Modified sentence: In 1563, Norwegian troops intercept the Swedish further at Elverum, which provide a strategic sharpen since it lie in on both north-south and east-west trade and travel road.

5.3.4 Next Steps

Are there other data augmentation techniques we can try which cover a larger subset of the sample and which can aid in balancing the label distribution?

Another idea is to displace constituents. For instance, if both quantified constituents follow the verb (such as two direct objects, or a direct object and an indirect object), can we switch them? We intend on exploring this in the future. However, given the lack of success of our automatic data augmentation strategies (other than passivization), we will not prioritize this as a potential avenue forward. Instead, we focus on collecting human annotations for a dataset and the development of a new model for the QSD task.

6 Model Development

While we have been almost exclusively concentrated on the construction of a large dataset, we also consider how an effective model can be built to tackle QSD. We say a few words on this topic here, and we intend on expanding on this in the fall term.

We strongly maintain that any successful model for this task must make use of world knowledge. Thus, it is essential that we expand on the work of Saba and Corriveau (2001) and Srinivasan and Yates (2009). NLP has advanced significantly since the appearance of these works and can likely provide more insight into how quantificational constraints can be learned from data. While we do not currently have a specific idea, we suspect this will involve collecting statistics from a large corpus and/or making use of word embeddings.

6.1 Learning Quantificational Constraints

Saba and Corriveau (2001) propose the formalism of *quantificational constraints*. A quantificational constraint takes two concepts C_1, C_2 and a relation R, and outputs the expected sizes m_1, m_2 of the

[2] `https://maelfabien.github.io/machinelearning/NLP_8/#synonym-replacement-sr`

sets C_1 and C_2 when they enter into relation R. For example, given $C_1 =$ "*doctor*", $C_2 =$ "*city*", and $R =$ "*livesIn*", we have that $QC(livesIn, doctor, city) = \langle 1+, 1 \rangle$. This means that at least one doctor lives in a city, but that it is implausible for the same doctor to live in many cities.

Srinivasan and Yates (2009) attempt to learn these quantificational constraints from data. They scrape the Web1Tgram corpus for n-grams which contain numerical quantifiers (e.g. "hundreds of" students) to estimate the probability that a concept has a particular set size. However, this method is relation-agnostic. That is, it is not concerned with the relations that the quantified entities in the n-gram enter into. They also collect data for relations between quantified entities (e.g. "she visited four countries"). However, these extractions are limited, and it is likely that a significant proportion of (C_1, R, C_2) triples are missing from this data.

Instead of the above approach, we propose using a language model to compute quantificational constraints. Though we initially considered BERT (Devlin et al., 2019), we realized that this is unsuitable for scoring the probability of phrases due its bidirectional nature. So, we turn to a unidirectional Transformer based language model, OpenAI GPT-2 (Radford et al., 2019). We ask it to provide the perplexity for manually curated quantificational constraints. For example, given the ambiguous sentence "A doctor lives in every city", we ask GPT-2 to give the perplexity score for the following quantificational constraints: "one doctor lives in one city", "one doctor lives in many cities". Sure enough, GPT-2 assigns a higher score to "one doctor lives in one city". This suggests that we should select the SECOND interpretation for the ambiguous sentence. Similarly, GPT-2 successfully predicts P(one diplomat occupies one chair) $>$ P(one diplomat occupies many chairs), and P(one official assigned to one court) $>$ P(one official assigned to many courts).

Now, our concern is whether or not such an approach generalizes to a larger collection of instances of quantifier scope disambiguation. Manual inspection of existing datasets and our sample from Section 5 tells us that this concern is justified.

As a preliminary test of our idea, we test GPT-2 on the dataset of Higgins and Sadock (2003). We work only with the universal quantifiers "all", "any", "both", "each", and "every", and the existential quantifiers "a", "some", "an", "either". As always, we only consider sentences with one universal and one existential quantifier. This leaves us with 32 sentences. We then manually modify the sentences by replacing the quantifiers in question with less ambiguous quantifiers, as we discussed previously with the example "A doctor lives in every city". We provide two substitutions per sentence: one aligned with the FIRST reading and the other aligned with the SECOND reading. For example, in the case of "A doctor lives in every city", the FIRST reading would correspond to "One doctor lives in many cities" and the SECOND reading would correspond to "One doctor lives in one city".

Of the 32 sentences we extracted, 26 have label FIRST and 6 have label SECOND. The universal quantifier takes wide scope in 13 out of the 32 instances and the existential quantifier takes wide scope in the other 19 instances.

Unfortunately, most of the extracted sentences do not fit the mold of our idea as nicely as the example with doctors and cities. As an illustration, consider the following sentence that was extracted: "Behind all the hoopla is some heavy-duty competition." We can infer that the second quantifier takes wide scope. But, how do we fit this into our framework of quantifier substitution? For the FIRST reading, we tried "Behind much hoopla is much heavy-duty competition". For the SECOND reading, we used "Behind much hoopla is one heavy-duty competition". This is unsatisfactory and arguably makes the problem harder to solve. If the two readings were provided to humans (without the original sentence), which one would a human say is more plausible?

We also found that there were several sentences which started with the phrase “According to some experts”, where “some” is viewed as an existential quantifier. In these cases, the “some” was always taken to have wide scope.

GPT-2 is 59.4% accurate on our collection of 32 sentences. While this is solid given the difficulty of the task, it is considerably worse than the baseline of taking the surface scope reading (81.3% accuracy). However, this likely has less to do with our use of GPT-2 and more to do with the way that we are posing the problem to GPT-2.

As a first small alteration to our approach. We provide the original ambiguous sentence followed by our constructed sentence with substituted quantifiers to GPT-2. Previously, we only provided the constructed sentence.

This approach proves to be even less successful than the first, obtaining only 46.9% accuracy, which is worse than random.

References

Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*, volume 16, pages 15–28, 2012.

Galen Andrew and Bill MacCartney. Statistical resolution of scope ambiguity in natural language. *Unpublished manuscript*, 2004.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2029. URL `https://www.aclweb.org/anthology/D18-2029`.

Emmanuel Chemla and Lewis Bott. Using structural priming to study scopal representations and operations. *Linguistic Inquiry*, 46(1):157–172, 2015.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Veena D Dwivedi. Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*, 8(11):e81461, 2013.

Roman Feiman, Mora Maldonado, and Jesse Snedeker. Priming quantifier scope: Reexamining the evidence against scope inversion. *Glossa: a journal of general linguistics*, 5(1), 2020.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. English gigaword third edition ldc2007t07. *Web Download. Philadelphia: Linguistic Data Consortium*, 2007.

Derrick Higgins and Jerrold M Sadock. A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96, 2003.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.768. URL https://www.aclweb.org/anthology/2020.acl-main.768.

Mehdi Manshadi and James Allen. Unrestricted quantifier scope disambiguation. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 51–59, 2011.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. 1993.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, 2019.

George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

Edward Newell, Kian Kenyon-Dean, and Jackie Chi Kit Cheung. Deconstructing and reconstructing word embedding algorithms. *arXiv preprint arXiv:1911.13280*, 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

Walid S Saba and Jean-Pierre Corriveau. Plausible reasoning and the resolution of quantifier scope ambiguities. *Studia Logica*, 67(2):271–289, 2001.

Prakash Srinivasan and Alexander Yates. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1465–1474, 2009.

Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6383–6389, 2019.

Adina Williams, Nikita Nangia, and Samuel Bowman. 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, 2018.