To What Extent Do Natural Language Understanding Datasets Correlate to Logical Reasoning? A Method for Diagnosing Logical Reasoning.

Yitian Li\textsuperscript{1,2†}, Jidong Tian\textsuperscript{1,2†}, Wenqing Chen\textsuperscript{1,2}, Caoyun Fan\textsuperscript{1,2}, Hao He\textsuperscript{1,2‡} and Yaohui Jin\textsuperscript{1,2}

\textsuperscript{1}MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
\textsuperscript{2}State Key Lab of Advanced Optical Communication System and Network, Shanghai Jiao Tong University
\{yitianli, frank92, wenqingchen, fcy3649, hehao, jinyh\}@sjtu.edu.cn

Abstract

Reasoning and knowledge-related skills are considered as two fundamental skills for natural language understanding (NLU) tasks such as machine reading comprehension (MRC) and natural language inference (NLI). However, it is not clear to what extent an NLU task defined on a dataset correlates to a specific NLU skill. On the one hand, evaluating the correlation requires an understanding of the significance of the NLU skill in a dataset. Significance judges whether a dataset includes sufficient material to help the model master this skill. On the other hand, it is also necessary to evaluate the dependence of the task on the NLU skill. Dependence is a measure of how much the task defined on a dataset depends on the skill. In this paper, we propose a systematic method to diagnose the correlations between an NLU dataset and a specific NLU skill, and then take a fundamental reasoning skill, logical reasoning, as an example for analysis. The method adopts a qualitative indicator to indicate the significance while adopting a quantitative indicator to measure the dependence. We perform diagnosis on 8 MRC datasets (including two types) and 3 NLI datasets and acquire intuitively reasonable results. We then perform the analysis to further understand the results and the proposed indicators. Based on the analysis, although the diagnostic method has some limitations, it is still an effective method to perform a basic diagnosis of the correlation between the dataset and logical reasoning skill, which also can be generalized to other NLU skills.

1 Introduction

Machine reading comprehension (MRC) and natural language inference (NLI) are used to benchmark natural language understanding (NLU) capabilities. Although a large number of NLU-related datasets have been proposed (Yang et al., 2018; Yatskar, 2019), it is hard to evaluate correlations between the dataset and NLU-related skills due to the lack of benchmark method (Richardson et al., 2020), which may create an obstacle to choose appropriate dataset for specific skill training. According to a widely accepted discourse comprehension theory, construction-integration model (Kintsch, 1991), there are two processes for humans to understand language: 1) understanding the concepts and discourses in the text and building their relationships with the real world; 2) synthesizing these concepts and discourses to form a consistent understanding. These two processes mainly correspond to two types of skills in NLU: Knowledge-related skills and reasoning skills (Baral et al., 2020; Boratko et al., 2018; Tian et al., 2021). Previous researches concentrate on evaluating the correlation between datasets and knowledge, such as assessing what types of knowledge are required to complete an NLU task (Sap et al., 2019a; Rogers et al., 2020) and offering various knowledge supplements based on those needs (Feng et al., 2020). However, compared to knowledge-based characteristics, reasoning abilities are more difficult to identify and quantify directly. Besides, some existing work using probes (supervised models trained to predict properties) mainly focuses on linguistic tasks (Hewitt and Liang, 2019; Conneau et al., 2018) and few studies diagnose the logical discrepancies between datasets.

Based on different deduction methods, reasoning skills can be divided into logical reasoning, co-referential reasoning, numerical reasoning, causal reasoning, etc (Sugawara et al., 2017). Among all these reasoning skills, logical reasoning is a fundamental skill widely used to understand natural language (Bhagavatula et al., 2020). In this paper, we take logical reasoning as an example and construct a systematic method to analyze the correlations between NLU datasets and a specific
reasoning skill. The correlations include two aspects: significance and dependence. 1) significance aims to judge whether a dataset includes sufficient (explicit or implicit) logical expressions; 2) dependence is used to measure how much the task on a dataset depends on the logical reasoning skill. To aid the diagnosis, a logical probe, consisting of a probe model and a probe dataset, is introduced. We select 8 MRC datasets (including two types) and 3 NLI datasets to perform diagnosis. We create a qualitative and quantitative indicator to reflect the association between the dataset and logical reasoning after training the probe model on various datasets in various ways. The results show: 1) Most NLI datasets are relatively strongly correlated to logical reasoning. 2) As for the comprehensive MRC datasets (Type1), the correlations are moderate which means that logical reasoning is not the only dominant reasoning skill in this type; 3) The correlations of different MRC datasets for specific anticipations (Type2) vary remarkably according to their purposes of design. Further cause analysis, which is consistent with the results, confirms the rationality of our method to a certain degree. In conclusion, this method offers a reasonable view of exploring the correlations between datasets and logical reasoning.

Our contributions are as follows:

- We propose a systematic method to diagnose the correlations between NLU datasets and reasoning skills and take logical reasoning as an example to validate the effectiveness of this method.
- In particular, two indicators are introduced to evaluate the correlation between the NLU dataset and reasoning skill.
- We conduct extensive experiments on 11 NLU datasets from both qualitative and quantitative analyses. Results show that Winogrande is the only dataset unable to judge the significance based on qualitative analysis, while QASC, SNLI and MNLI show relatively high dependence on logical reasoning based on quantitative analysis.

2 Related Work

NLU Datasets. Recently, the number of NLU datasets has exponentially increased (Zeng et al., 2020). Such datasets mainly include MRC datasets and NLI datasets. MRC datasets, such as SQuAD (Wang et al., 2016), CoQA (Reddy et al., 2019) and DROP (Dua et al., 2019) are designed to test whether machines can answer the text-related questions or not, while NLI datasets, such as SNLI (Bowman et al., 2015) and MNLI (Nangia et al., 2017), are constructed to explore whether models can detect inferential relationships between natural language descriptions (Richardson et al., 2020) or not. Overall, all these datasets aim to define tasks to evaluate whether machines can understand the natural language as humans do.

Analysis for NLU. Recently, more and more researches focus on what models have really learned in NLU tasks. In this field, some researches attempt to understand the knowledge (Ghosal et al., 2021; Fang et al., 2021) and reasoning skills (Richardson et al., 2020) in language models, while others mainly focus on systematic evaluations of NLU models (Tenney et al., 2019; Ribeiro et al., 2020). These studies have provided foundations to further understand NLU. In addition, some researchers begin to pay attention to the analysis of NLU datasets (Baradaran et al., 2020). For example, Sugawara et al. (2020) have provided an ablation-based method to understand the tasks defined on the NLU datasets. However, such researches are still rare and further investigations are required.

3 Methodology

In this paper, we limit our investigation to a fundamental form of logical reasoning: conjunctive implications with negation (Musen and Lei, 1988) which uses multiple conditions to derive the final conclusion and can be expressed as: \[((-p_1, (neg)p_2, \cdots, (neg)p_n) \rightarrow q, \] where \(p_i, (i = 0, 1, \cdots, n)\) represents each condition and \(q\) is the conclusion. Although logical reasoning has many complex forms, conjunctive implication with negation is the most commonly used one in daily conversation and can cover most situations in natural language (Allwood et al., 1977).

Our method adopts control variates to control what models learn. Supported by a logical probe, we train models in the different manners on the diverse datasets to adjust the skills mastered by models. Based on the method, we design a qualitative indicator and a quantitative indicator to indicate significance and dependence, respectively. Next, we will firstly introduce the logical probe and then introduce two indicators and two corresponding
processes.

3.1 Logical Probe

Logical probe includes a probe dataset and a probe model. The probe dataset has two main functions: 1) to train the probe model and enable it to master logical reasoning without increasing extra knowledge; 2) to test whether a model has mastered logical reasoning or not. Therefore, the probe dataset is required to meet the following three conditions:

- The dataset will involve as little explicit and implicit knowledge as possible, while expressions in the dataset should conform to natural language form.
- It should contain a large number of logical expressions and the task defined on the dataset should relate strongly to logical reasoning.

As for the probe model, it is the original model that will be trained on the different datasets in diverse manners. Therefore, the probe model is required to meet the following three conditions:

- The probe model itself can be used as a knowledge base.
- The knowledge stored in the model’s weights can be updated after being trained in a specific manner.
- The model can master logical reasoning after being trained on the probe dataset.

3.2 Qualitative Diagnosis

A qualitative process is designed to diagnose the significance of a dataset to logical reasoning which indicates whether a dataset includes sufficient logical materials to enable the model to master logical reasoning. To perform qualitative diagnosis, we firstly involve a presupposition that as long as there exists one model that can master logical reasoning to a certain degree through the dataset to be diagnosed, we can assert that the dataset is significant to logical reasoning. According to the presupposition, the qualitative process includes two steps shown in Fig 1 and Alg 1. The first step is to train the probe model \( M_0 \) on the dataset \( D \) to be diagnosed in the supervised manner. Then we acquire a trained model \( M_1 \) which has acquired the dominant patterns in \( D \). The second step is to test \( M_1 \) on the probe dataset \( P \) to acquire the metric \( R_1 \).

![Figure 1: Qualitative process.](image)

Note that in the testing step, we re-train the fully-connected classifier with fixed \( M_1 \) before testing. The qualitative indicator can be calculated by the Eq 1.

\[
I_1 = \text{sign}(R_1, \delta) = \begin{cases} +, & R_1 \geq \delta \\ - & R_1 < \delta \end{cases}
\]  

(1)

Where “+” means “significant to the logical reasoning” while “−” means “unable to judge” (the qualitative diagnosis has a natural limitation in which we cannot traverse all models to fully verify the existential presupposition). \( \delta \) is the threshold to judge the significance (Since the probability of random selection for the binary classification problem is 0.5, we can approximate that \( R_1 \) is equivalent to the result of random selection if \( R_1 < \delta = 0.55 \)).

**Algorithm 1 Qualitative Process.**

Require: NLU datasets \( D = [D_1, D_2, \cdots, D_n] \);
Prove dataset \( P \); Probe Model \( M_0 \)

Ensure: Qualitative Indicator, \( I_1 = [I_1^1, I_1^2, \cdots, I_1^n] \)

1: Initialize \( M_0 \) with pre-trained parameters;
2: Initialize \( I_1 = [] \)
3: for \( i = 1 \) to \( n \) do
4: \( M_1^i = \text{SupervisedTrain}(M_0, D_i) \)
5: \( R_1^i = \text{Test}(M_1^i, P) \)
6: \( I_1^i = \text{Sign}(0, R_1^i - \delta) \)
7: \( I_1, \text{Append}(I_1^i) \)
8: end for
9: return \( I_1 \);

3.3 Quantitative Diagnosis

A quantitative process is designed to diagnose the dependence of a dataset on logical reasoning, which aims to answer how much a task defined on the dataset depends on the logical reasoning skill. Here we make a hypothesis that if the only difference between the two models is whether logical reasoning has been mastered, the gap of the
We select 11 NLU datasets to perform diagnosis. These datasets can be classified into three types: NLI datasets and two types of MRC datasets. The NLI datasets include three commonly used ones: SNLI (Bowman et al., 2015), MNLI (Nangia et al., 2017) and αNLI (Bhagavatula et al., 2020). Type 1 MRC datasets (comprehensive MRC datasets) include BoolQ (Clark et al., 2019), DROP (Dua et al., 2019) and CODAH (Chen et al., 2019), all with tasks requiring diverse NLU skills to solve. Type 2 MRC datasets (specific MRC datasets) are constructed to benchmark one dominant NLU skill, including QASC (Khot et al., 2020), ReCLor (Yu et al., 2020), SocialIQA (Sap et al., 2019b), QuaRTz (Tafjord et al., 2019) and Winogrande (Sakaguchi et al., 2020).

In order to make the diagnosis results on different datasets comparable, we process the NLI and MRC datasets separately. Given an item \( D = \{\text{context}, \text{question}, \text{answer}\} \) in an MRC dataset, we combine question and answer to a statement and transfer the \( D \) to a binary classification represented by \( D' = \{\text{context}, \text{statement}, \text{label}\} \), where label is True or False.

As for the item \( D = \{\text{fact}1, \text{fact}2, \text{label}\} \) in an NLI dataset, we just set fact1 as context and fact2 as statement, and change label \( \in \{\text{entailment}, \text{neural}, \text{contradiction}\} \) to label \( \in \{\text{True}, \text{False}\} \) (entailment is True; neural and contradiction are False based on the closed world assumption, CWA). To ensure the balance of samples, we adjust the number of negative samples by removing or constructing. The statistics of the processed datasets are shown in Table 2.

4 Diagnosis

4.1 NLU Datasets to Be Diagnosed

We select 11 NLU datasets to perform diagnosis. The quantitative process is shown in the Fig 2 and Alg 2. The test algorithm in both Alg 1 and Alg 2 is shown in Alg 3.

Algorithm 2 Quantitative Process.

Require: Diagnosed datasets \( D = \{D_1, D_2, \cdots, D_n\} \); Probing dataset \( P \); Probing Model \( M_0 \)

Ensure: Quantitative Indicator, \( I_2 = [I_{21}, I_{22}, \cdots, I_{2n}] \)

1: Initialize \( M_0 \) with pre-trained parameters;
2: Initialize \( I_2 = [\] \)
3: for \( i = 1 \) to \( n \) do
4: \( M'_2 = \text{UnsupervisedTraining}(M_0, D_i) \)
5: \( R'_2 = \text{Testing}(M'_2, D_i) \)
6: \( M'_3 = \text{SupervisedTraining}(M'_2, P) \)
7: \( R'_3 = \text{Testing}(M'_3, D_i) \)
8: \( I_i^2 = \frac{R'_3 - R'_2}{R'_3} \times 100 \)
9: \( I_2. \text{Append}(I_i^2) \)
10: end for
11: return \( I_2; \)

Algorithm 3 Test (including Re-train FC Classifier).

Require: Model \( M \); Dataset \( D \)

Ensure: Metric \( R \)

1: Initialize fully-connected classifier \( C \) randomly;
2: Fix \( M \)
3: \( C' = \text{SupervisedTrain}([M, C], D) \)
4: \( R = \text{CalMetric}([M, C'], D) \)
5: return \( R; \)
We trained our model on NVIDIA 16GB Tesla P100 and V100 GPUs. For fine-tuning BERT, the model costs around 22GB memory with the batch size of 16. The hyper-parameters are shown in Table 1.

| Paras.                  | Unsupervised | Supervised | Training |
|-------------------------|--------------|------------|----------|
| batch size              | 16           | 16         | 16       |
| lr                      | $-10^{-5}$   | $10^{-6}$  | $10^{-4}$|
| lr for BERT             | $5\times10^{-6}$ | $5\times10^{-6}$ | $10^{-5}$ |
| decay rate              | 0.9          | 0.9        | 0.8      |
| l2 coeff.               | $10^{-5}$    | $10^{-5}$  | $10^{-5}$|
| early stop              | 5            | 5          | 5        |
| epochs                  | 20           | 20         | 20       |
| optimizer               | ADAMW        | ADAMW      | ADAMW    |

Table 1: Hyper-parameter settings.

| Datasets | Average No. Data (k) |
|----------|----------------------|
|          | Length | Train | Dev. | Test |
| SNLI     | 21     | 300.0 | 6.0  | 6.0  |
| MNLI     | 30     | 245.0 | 13.5 | 13.4 |
| αNLI     | 33     | 339.0 | 3.0  | 6.0  |
| BoolQ    | 102    | 4.6   | 2.4  | 2.4  |
| DROP     | 197    | 132.0 | 18.5 | 14.7 |
| CODAH    | 17     | 3.5   | 1.0  | 1.0  |
| QASC     | 108    | 13.9  | 1.9  | 2.4  |
| ReCLor   | 97     | 8.2   | 1.0  | 1.0  |
| SocialIQA| 23     | 33.4  | 2.0  | 2.2  |
| QuaRTz   | 37     | 5.3   | 0.7  | 1.3  |
| Winogrande| 20    | 78.3  | 2.5  | 2.5  |

| Probe Dataset | 103 | 70.1 | 9.9 | 20.0 |

Table 2: Statistics of NLU datasets to be diagnosed and the probe dataset.

4.2 Probe Dataset

In the diagnosis, we use the dataset proposed by Clark et al. (2020) as the probe dataset (An example of which is shown in Fig 3) for it clearly satisfying the last two conditions. Based on the construction rules (constructing propositions through meaningless sentences) of the probe dataset, it will not involve explicit knowledge into models. To validate that the implicit knowledge involved by the dataset can be negligible, we also make a simple statistical analysis in Preliminary 1. Therefore, this dataset is suitable to be the probe dataset. In practice, we use the 3-hop dataset to train the probe model because Clark et al. (2020) have illustrated that Transformers trained on this dataset has good generalization capability. Statistical information of the probe dataset is shown in Table 2.

4.3 Probe Model

BERT (Devlin et al., 2019) naturally meets all three conditions of the probe model. Firstly, BERT itself can be used as the knowledge base since it contains a large amount of knowledge and its architecture has the ability to update knowledge after being trained in the unsupervised manner (the same way to train the language model) (Rogers et al., 2020; Petroni et al., 2019). Secondly, it has been proved that Transformers, the basic architecture of BERT, can master the generalizable logical reasoning (Clark et al., 2020; Hahn et al., 2020). Therefore, we select BERT as the probe model. Although BERT can satisfy all three conditions, we still need to prove that no matter what initial weights are, the architecture of Transformers always has the ability to master generalizable logical reasoning. These two preliminaries are shown in Preliminary 2 and Preliminary 3, respectively.

4.4 Preliminaries

Preliminary 1: Validating that implicit knowledge involved by the probe dataset can be negligible. Implicit knowledge is always hidden in the representative vectors of words. We have made a simple statistics that the vocabulary size of the probe dataset is 67. This means that the proportion of overlapping words to the NLU dataset’s vocabulary is very low, almost negligible, which is intuitive evidence that implicit knowledge involved by the probe dataset can be negligible.

Preliminary 2: Proving that BERT will not master logical reasoning after being trained in the unsupervised manner. We conduct a controlled experiment to compare three kinds of models trained and tested on the probe dataset. We use...
BERT + fully-connected classifier as the original architectures. For the first model, we fix BERT and only train the fully-connected classifier. For the second model, we firstly train BERT in the unsupervised manner, and then fix it and only train the fully-connected classifier. For the third model, we fine-tune BERT and train the classifier at the same time. The final results of the first two models are almost equivalent to random selection (Accuracy: 50.41% and 50.00%, respectively) but the result of the third model can reach 98.69%. This has validated that BERT cannot master logical reasoning after being trained in the unsupervised manner.

Preliminary 3: Validating that Transformers always has the ability to master logical reasoning no matter what initial weights are. We perform this validation based on the process shown in Fig 4. Firstly, we acquire $M_3$ following the quantitative process. Secondly, we test $M_3$ on the probe dataset. We compare the results $R_{V,3}$ with the original result in (Clark et al., 2020), which is shown in Fig 5 (dotted line represents the original result). The comparison illustrates that no matter what initial weights are, the performance will finally reach a similar level to the original result’s level.

5 Results and Analysis

5.1 Overall Results of Diagnosis

Based on the diagnosis, we acquire a qualitative indicator $I_1$ and a quantitative indicator $I_2$ for each NLU to be diagnosed and the results are shown in Table 3. Based on the results, we firstly answer two questions related to the correlation on logical reasoning.

1. Whether does a dataset have sufficient explicit or implicit logical expressions that enables models to master logical reasoning or not? Based on results ($I_1$) of the qualitative diagnosis, almost all datasets (except for Winogradne) shows significance to logical reasoning, which means they have a certain amount of logical reasoning to enable the probe model to master logical reasoning. Winogradne, the only exception that cannot be judged by the probe model, will be further analyzed in the section of Exception Analysis. To compare NLI and MRC datasets, we adopt another threshold $\delta' = 0.6$ to distinguish “very significant” (“++”) and “significant” (“+”). Generally, NLI datasets are more significant than MRC datasets. We speculate that reasons for the conclusion are: 1) Logical patterns in the NLI datasets are more likely to be captured as the contexts in these datasets are much simpler; 2) MRC datasets usually contain complicated contexts requiring a variety of reasoning skills, which makes it hard to expose the logical patterns.

2. How much does the task defined on a dataset depend on the logical reasoning skill? Although most NLU datasets can provide materials of logical reasoning, not all tasks defined on these datasets have strong dependences on logical reasoning. We have shown the dependence indicator $I_2$ in Fig.6. From the figure, we can find that NLI tasks, except for $\alpha$NLI, have relatively high dependences on logical reasoning. This means that NLI tasks often highly rely on logical reasoning, which plays a dominant role in these tasks. These results are roughly consistent with the definition and purpose of the NLI tasks (The outlier in NLI datasets, $\alpha$NLI, will be further analyzed in the section of Exception Analysis). Among comprehensive MRC tasks (Type 1), we can find the dependence indicators for all three tasks are moderate, which may be due to their requirements of multiple NLU skills. In terms of MRC datasets for specific purposes (Type 2), the results show a remarkable difference among these tasks. On the
Table 3: Diagnosed Results: $I_1$ and $I_2$ are two indicators for the qualitative process and the quantitative process, respectively. The threshold $\delta$ of qualitative indicator is 55%. To compare NLI and MRC datasets, we adopt another threshold $\delta' = 60\%$ to distinguish “very significant (++)” and “significant (+)”. Therefore, “-” means “unable to judge”, “+” means “significant” and “++” means “very significant”.

5.2 Exception Analysis

In this section, we conduct a detailed analysis of the exceptions, $\alpha$NLI, ReCLor and Winogrande, mentioned above to further understand the causes of the results. Meanwhile, on the basis of the analysis, we also acquire a better understanding of the limitation of the proposed quantitative indicator $I_2$.

1. Winogrande. From the qualitative indicator $I_1$, Winogrande is the only dataset unable to judge whether it contains sufficient logical expressions or not. We take a representative pair of examples in the dataset as the case to analyze. The case includes (Context: The trophy does not fit into the suitcase, Statement: because trophy is too large, Label: True) and (Context: The trophy does not fit into the suitcase, Statement: because suitcase is too large, Label: False). It is obvious that no logical expressions or underlying logical forms.

2. $\alpha$NLI. $\alpha$NLI seems an outlier which has a lower $I_2$ than other NLI tasks. Based on our analysis, we find that although $\alpha$NLI is classified into the NLI category, it is designed to benchmark sequential reasoning in a narrative text rather than logical reasoning. An example of $\alpha$NLI is (Context: Jill had a pet cat. She was able to remove the fleas by sprinkling salt on her floors, Statement: The cat had fleas, Label: True). It is obvious that the example requires the model to be able to understand the sequential relation between context and statement rather than apply logical reasoning. Therefore, the dependence

| Diagnosis | Acc. (%) | Ind. | NLI | MRC(Type1) | MRC(Type2) |
|-----------|---------|------|-----|------------|------------|
| Qualitative | $R_1$ | 60.35 | 61.32 | 57.12 | 58.14 | 58.38 | 59.13 | 58.39 | 58.21 | 60.35 | 57.30 | 50.77 |
| Quantitative | $R_2$ | 50.29 | 51.11 | 56.80 | 55.40 | 51.60 | 57.90 | 51.56 | 51.34 |
| Indicators | $I_1$ | ++ | ++ | + | + | ++ | + | ++ | + | - |
| $I_2$ | 10.31 | 9.13 | 4.36 | 4.30 | 5.00 | 3.33 | 14.35 | 2.27 | 3.63 | 1.26 | 2.61 |

Figure 6: Quantitative indicators. Among all three kinds of datasets, most NLI datasets (except for $\alpha$NLI) have high indicators. The indicators of comprehensive MRC datasets are relatively moderate. Specific MRC datasets contain QASC with the highest $I_2$ and QuaTRz with the lowest $I_2$, which are complicated.

one hand, tasks such as QASC ($I_1 = 14.35$) show a strong dependence on logical reasoning. On the other hand, tasks such as SocialIQA ($I_2 = 3.63$), QuaTRz ($I_2 = 1.26$) and Winogrande ($I_2 = 2.61$), show almost no dependence on logical reasoning. Based on the reports from original papers, the corresponding dominant skills for four MRC datasets (Type 2) are logical reasoning (Khot et al., 2020), knowledge-based skill (Sap et al., 2019b), numerical reasoning (Tafjord et al., 2019) and coreferential reasoning (Sakaguchi et al., 2020), respectively. This is consistent with $I_2$ indicating that only QASC has a strong dependence on logical reasoning among these four datasets. Surprisingly, although ReCLor, a Type 2 MRC dataset, is designed for the purpose of evaluating logical reasoning (Yu et al., 2020), $I_2$ cannot indicate that the dataset has a strong dependence on logical reasoning. Aiming at this exception, we will also perform individual analysis in the section of Exception Analysis.

5.2 Exception Analysis

In this section, we conduct a detailed analysis of the exceptions, $\alpha$NLI, ReCLor and Winogrande, mentioned above to further understand the causes of the results. Meanwhile, on the basis of the analysis, we also acquire a better understanding of the limitation of the proposed quantitative indicator $I_2$.

1. Winogrande. From the qualitative indicator $I_1$, Winogrande is the only dataset unable to judge whether it contains sufficient logical expressions or not. We take a representative pair of examples in the dataset as the case to analyze. The case includes (Context: The trophy does not fit into the suitcase, Statement: because trophy is too large, Label: True) and (Context: The trophy does not fit into the suitcase, Statement: because suitcase is too large, Label: False). It is obvious that no logical expressions or underlying logical forms.

2. $\alpha$NLI. $\alpha$NLI seems an outlier which has a lower $I_2$ than other NLI tasks. Based on our analysis, we find that although $\alpha$NLI is classified into the NLI category, it is designed to benchmark sequential reasoning in a narrative text rather than logical reasoning. An example of $\alpha$NLI is (Context: Jill had a pet cat. She was able to remove the fleas by sprinkling salt on her floors, Statement: The cat had fleas, Label: True). It is obvious that the example requires the model to be able to understand the sequential relation between context and statement rather than apply logical reasoning. Therefore, the dependence
Figure 7: Line charts of quantitative indicator $I_2$ and paired difference $\Delta$. The results show that $I_2$ is significantly correlated to $\Delta$.

of $\alpha$NLI on logical reasoning is weak.

3. ReCLor. As Yu et al. (2020) reported, ReCLor is constructed from GMAT and LAST to evaluate logical reasoning. However, in our diagnosis, the indicator of ReCLor is very low. Based on our analysis, the causes may include: 1) To solve the task defined on ReCLor requires a large amount of legal-related domain knowledge and commonsense, which is not sufficiently acquired by the probe model. Therefore, the model cannot capture the underlying logical forms; 2) Although logical reasoning is the dominant skill in ReCLor, it cannot be decoupled from multiple complex skills. Therefore, we can further understand the limitations of our proposed method from the analysis of ReCLor. Actually, $I_2$ reflects the lower bound of the dependence on logical reasoning, as it can just indicate the part satisfying the following conditions:

- logical reasoning is the dominant skill under the condition that knowledge acquired by the probe model is sufficient.
- logical features can be decoupled from the complex reasoning features.

Despite the limitations, $I_2$ is still a reasonable indicator to evaluate the dependence of NLU tasks on logical reasoning. Further illustration will be given in the next section.

5.3 Paired Analysis

Intuitively, models with logical reasoning are sensitive to subtle differences between logically true propositions and logically false propositions. Based on this intuition, we perform a paired analysis to provide intuitive evidence that $M_3$ does master logical reasoning compared with $M_2$. This is indirect evidence that our proposed indicator, $I_2$, is reasonable to indicate the dependence between the NLU dataset and logical reasoning. To perform the analysis, we first reconstruct paired sets comprising pairs of positive and negative samples on four datasets, SNLI (NLI), CODAH (Type 1 MRC), QASC (Type 2 MRC) and SocialIQA (Type 2 MRC). These four datasets have a common feature that for each positive example, we can extract a corresponding negative example directly from the datasets and vice versa. The only difference between the pair is that two samples have different keywords in their statements. This setting can ensure that if a sample includes a logically true or false proposition, the counterpart must include a logically false or true proposition. Next, we test $M_2$ and $M_3$ directly on paired sets and calculate their paired accuracies. Finally, we use the gap between $M_3$’s accuracy and $M_2$’s accuracy $\Delta$ to conduct the analysis. These results are shown in Table 4. From the table, $\Delta$ for each dataset is larger than 10%, which is evidence that the difference between $M_3$ and $M_2$ is decidedly due to logical reasoning.

| Dataset   | $M_3$ | $M_2$ | $\Delta$ |
|-----------|-------|-------|----------|
| SNLI      | 41.3  | 18.5  | 22.8     |
| CODAH     | 18.5  | 30.8  | 12.3     |
| QASC      | 11.6  | 33.6  | 22.0     |
| SocialIQA | 18.4  | 29.0  | 10.6     |

Table 4: Paired differences of SNLI, CODAH, QASC and SocialIQA.

Moreover, we further make a horizontal comparison between datasets and Fig 7 is the line charts of $I_2$ and $\Delta$. According to the analysis above, $\Delta$ is a reasonable indicator to measure the dependence on logical reasoning. From the figure, a significant positive correlation between $I_2$ and $\Delta$ exists. This is intuitive evidence that $I_2$ is also a reasonable indicator to reflect the dependence of a dataset on logical reasoning skill.

5.4 Case Study

In this part, we perform a case study to show the logical structure detected by the quantitative indicator $I_2$. We select a case from QASC (shown in Fig.8) which shows the strongest dependence on logical reasoning. From the case, we list the logical structure of the case for reasoning $p_1 \land p_2 \rightarrow q_1$. This is a typical logical form of the conjunctive implication which can be diagnosed by $I_2$. 

1715
context: Beads of water are formed by water vapor condensing ($p_1$). Clouds are made of water vapor ($p_2$). Condensation is the change of water vapor to a liquid ($p_3$). An example of water vapor is steam ($p_4$). Condensation of water vapor occurs during the chilling season ($p_5$).

statement: Beads of water is formed by clouds ($q_1$).

Liquid of water is formed by clouds ($q_2$).

label: True for $q_1$ and False for $q_2$.

logical structure for reasoning: $p_1 \land p_2 \rightarrow q_1$

Figure 8: A logical case from QASC

6 Conclusions and Future Work

In this work, we present a novel framework, which can diagnose the correlation between the NLU dataset and a specific skill and we probe a fundamental reasoning skill, logical reasoning, on 11 NLU datasets. Our framework involves a logical probe to conduct diagnosis and defines a qualitative process and a quantitative process to calculate two indicators. From the results, We observe that 1) Most NLI datasets have a relatively strong correlation with logical reasoning. 2) The correlations between Type 1 MRC datasets and logical reasoning are moderate because logical reasoning is not the only dominant skill in these datasets. 3) The dependences of Type 2 MRC datasets are not always exactly consistent with their intended purpose. Based on the analysis, although there are several limitations in our proposed method, this work is still a reasonable attempt to deeply understand the relationship between the dataset and a specific NLU skill. In future works, we will focus on: 1) exploring the solution to the limitations of the proposed method; 2) build associations for different datasets that require the same NLU capabilities.

Acknowledgements

This work was supported by the Shanghai Municipal Science and Technology Major Project (2021SHZDX0102), and the Shanghai Science and Technology Innovation Action Plan (20511102600).

References

Jens Allwood, Lars Andersson, and Östen Dahl. 1977. Logic in linguistics.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2020. A survey on machine reading comprehension systems. In CoRR.

Chitta Baral, Pratyay Banerjee, Kuntal Kumar Pal, and Arindam Mitra. 2020. Natural language QA approaches using reasoning with external knowledge. In CoRR.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Nahah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2020. Abductive commonsense reasoning. In ICLR.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In EMNLP.

Michael Chen, Mike D’Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: an adversarially authored question-answer dataset for common sense. In CoRR.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In NAACL-HLT.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. In IJCAI.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $\$&!!*$ vector: Probing sentence embeddings for linguistic properties. In ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In NAACL-HLT.
