ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 47, No. 3, pp. 298-309, 2020. 3
https://doi.org/10.5626/JOK.2020.47.3.298

# 기계 독해 질의 응답 모델의 한계점 파악을 위한 어휘적 정답유형 분석

## (Analysis of the Semantic Answer Types to Understand the Limitations of MRQA Models)

임 도 연 † (Doyeon Lim)　아리츠 푸에르토 산 로만 †† (Haritz Puerto San Roman)　맹 성 현 ††† (Sung-Hyon Myaeng)

**요 약** 최근 MRQA 모델들의 성능이 인간을 넘어섰다. 그리하여 MRQA 모델의 새로운 가능성들을 찾기 위해 새로운 데이터 셋들이 소개되고 있다. 하지만, 이전 MRQA모델들이 어떤 유형에서 문제를 잘 풀고 어떤 한계점이 있는지 자세한 분석을 통해 새로운 데이터셋을 제시하는 경우는 거의 없었다. 이 연구에서는 MRQA가 극복했다고 여겨지는 SQuAD 데이터 셋을 분석하여 MRQA가 언어를 이해한 것이 아니라 특정한 패턴을 찾아냈다는 것을 밝혀낸다. 이 과정에서 기존 QA데이터 셋에서 주로 등장하는 wh-word와 Lexical Answer Type (LAT) 정보에 많은 모델들이 특히 집중하고 있다는 것을 밝히고, 그 때문에 질의와 문서의 정보를 충분히 이해하지 못하고 있다는 것을 정성, 정량적인 수치로 보였다. 이러한 분석을 바탕으로 앞으로 MRQA의 데이터셋의 방향과 모델들이 극복해야할 한계점을 제시하였다.

**키워드:** 기계 독해 질의 응답, 질의 분석, 트랜스포머 언어 모델, 정답유형 분석

***Abstract*** Recently, the performance of Machine Reading Question Answering (MRQA) models has surpassed humans on datasets such as SQuAD. For further advances in MRQA techniques, new datasets are being introduced. However, they are rarely based on a deep understanding of the QA capabilities of the existing models tested on the previous datasets. In this study, we analyze the SQuAD dataset quantitatively and qualitatively to demonstrate how the MRQA models answer the questions. It turns out that the current MRQA models rely heavily on the use of wh-words and Lexical Answer Types (LAT) in the questions instead of using the meanings of the entire questions and the evidence documents. Based on this analysis, we present the directions for new datasets so that they can facilitate the advancement of current QA techniques centered around the MRQA models.

**Keywords:** machine reading question answering, query analysis, transformer language models, answer type

· This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT. (2017M3C4A7065962)
· 이 논문은 2019 한국컴퓨터종합학술대회에서 '최신 질의응답 독해모델의 정답 유형 활용능력에 대한 분석'의 제목으로 발표된 논문을 확장한 것임

† 비 회 원 : 한국과학기술원 전산학부 학생
dylim@kaist.ac.kr
†† 학생회원 : 한국과학기술원 전산학부 학생
haritzpuerto94@kaist.ac.kr
††† 종신회원 : 한국과학기술원 전산학부 교수(KAIST)
myaeng@kaist.ac.kr
(Corresponding author임)

논문접수 : 2019년 9월 24일
(Received 24 September 2019)
논문수정 : 2019년 12월 2일
(Revised 2 December 2019)
심사완료 : 2020년 1월 22일
(Accepted 22 January 2020)


정보과학회논문지 제47권 제3호(2020. 3)

## 1. Introduction

Recently, Machine Reading Question Answering (MRQA) has reached a new state of the art, owing to the advances in pre-trained language models based on Transformer architectures. Models based on BERT [1] and XLNet [2] have achieved a performance better than humans on the SQuAD dataset [3]. Solving this MRQA task requires reading and understanding a question and a document in which the answer appears explicitly. Current research in this field proposes new models for improvements for the existing QA tasks [4-6]. However, there are few studies on why in general these models can perform so well in these tasks. Understanding the true capabilities of the current models is indispensable to improve the performance of MRQA models. Now that these models perform better than humans on SQuAD, we need to look for more difficult QA tasks like HotPotQA [7] or Natural Questions [8]. In conjunction with these efforts, we need a better understanding of why MRQA has achieved this performance on existing QA datasets like SQuAD and what limitations exist.

In this study, we hypothesize that MRQA models do not attempt to "understand" the meaning of the documents and questions (queries), but exploit some simple patterns that are present in the datasets like SQuAD. These MRQA datasets usually have questions where the wh-word is followed by a lexical answer type (LAT). This LAT is revealed by an explicit word that appears on what- and which-questions and specifies the type of answer needed. For example, in the question 'What color was used to emphasize the 50th anniversary of the Super Bowl?', the LAT is color. Since this pattern is so frequent, the existing MRQA models only try to learn this pattern among others to predict the answers. In other words, these models look for the answer type of a question and look for entities in the document with the same entity type without deeply considering the meaning of the rest of the question and document. To demonstrate this, we perform a quantitative analysis of the errors made by four different recent QA models centered around the answer types (AT) of the questions in SQuAD. Using these analyses, we suggest several requisites that future MRQA datasets should include for developing better models.

## 2. Related Work

### 2.1 Limitation of MRQA Models

The idea that MRQA models may only learn how to answer questions using the answer type (AT) has already been suggested [4]. In that work, the authors propose two neural QA heuristics, one of which is based on AT. After doing qualitative analysis, the authors suggested that their models, and maybe other models, mostly learn how to match the answer type with the context. However, they lack an in-depth analysis that proves their conjectures. [9] performed for the first time an in-depth study of AT features for Machine Reading Question Answering (MRQA) models. In that work, the authors discovered that in SQuAD [3] most of the models have a high rate of errors in questions without AT. In this work, we propose to complement that study with further quantitative and qualitative experiments.

### 2.2 New Datasets

One of the problems of SQuAD v1.1 is that it provides questions and evidence documents that always contain answers, so the models only need to look for the most relevant span to the question, instead of attempting to decide that the span is the actual answer to the question. Thus, in SQuAD v2 questions without answers were proposed to make the QA systems more robust [10]. Another problem of SQuAD is that it is based on single-hop reasoning, i.e. to answer the questions, the model only needs to have the ability of reasoning within a single paragraph or document. Thus, recently a new dataset for multi-hop reasoning has been proposed [7]. In this new dataset, the models need to be able to reason across documents to answer the questions. This added requirement makes this dataset unique and more difficult. However, in all these works, the problems of the current MRQA datasets are not analyzed in sufficient depth. In this work, instead of proposing a dataset, we reveal the flaws and limitations by analyzing SQuAD, the most popular dataset for MRQA, and the state-of-the-art models on this dataset [2], to shed light on how answers are generated by existing MRQA models and to help

creating more challenging questions for advanced QA capabilities.

## 3. Analysis Methods

In this section, we explain the experiments we conducted to analyze the performance of MRQA in SQuAD [3].

### 3.1 Quantitative Analysis

We performed a quantitative analysis of the limitations of SQuAD using four recent MRQA models: BiDAF [5], DocumentQA [6], BERT [1] and XLNet [2]. We selected these four models because they represent the two most predominant families of MRQA models available at this point: transformer-based models and recent non-transformer based models. BiDAF is a well-known non-transformer based baseline for MRQA. DocumentQA is another non-transformer based model with high performance on datasets like SQuAD and TriviaQA [11]. BERT is a transformer-based model that established a new state of the art on most of the GLUE tasks, including SQuAD. Lastly, XLNet is the newest released transformer-based model that established a new state of the art on SQuAD among other tasks.

Several experiments were conducted to understand how MRQA models can solve questions in the datasets like SQuAD. First, we analyzed which wh-questions are more difficult to answer by counting the number of incorrectly answered questions per wh-type for the four models. Similarly, for all what- and which-questions, we analyzed the most difficult LAT by counting the number of incorrectly answered questions for each type across the four models.

Second, we checked the exitence of a correlation between the frequency of the LAT in the training set is correlated to the performance of the MRQA models to solve questions including those LATs. Lastly, we obtained statistics of the LATs that the MRQA models must solve to improve its overall performance. In order to do that, we designed a measure called urgency score to see in which LATs the MRQA models have problems and as a result to show their weaknesses. This score $U_k$ captures the failure rate of LAT k and its frequency on the dev set such that the higher the failure rate and the frequency, the higher the score is.

$$U_k = F_k \times log(frequency\ of\ k\ in\ dev\ set)$$
$$F_k = \sum_{n=2}^{4} \frac{1}{2}^{(4-n)} \frac{wrong(n,k)}{total(k)} \quad (1)$$

where $k \in L$ and $L$ is the set of all LATs in the dataset. If a certain LAT appears frequently in the dataset and has a high rate of being determined incorrectly, new MRQA models should try to perform better on it to maximize the increase of its performance with respect to previous models. We use the

Table 1 Failure case categories

| | | Question | |
|---|---|---|---|
| **Failure case** | **Explanation** | **Golden answer** | **Predicted answer** |
| Alias | The model predicted the right answer but it is not included in the list of golden answers. Usually, the predicted answer is very similar to one of the golden answers. | What kind of rail system is Metro Trains Melbourne? | |
| | | passenger | passenger system |
| Boundary | The model cannot return a full answer or the answer has non-necessary information | What leads to lower income inequality? | |
| | | redistribution mechanisms such as social welfare programs | social welfare programs |
| Wrong context | The context of the sentence with the predicted answer does not match the context of the question. | What was the name of the first Huguenot church in the New World? | |
| | | L'Église française à la Nouvelle-Amsterdam | L'Eglise du Saint-Esprit |
| Wrong answer type | The entity type of the answer does not match the expected answer type of the question. | What is the name of country which has Washington, D.C. as capital? | |
| | | the United States | Obama |
| Modifier | The model doesn't include an essential modifier in the answer. It is common in *how many* questions. | When was Kublai's administration running out of money? | |
| | | after 1279 | 1279 |

logarithm of the number of occurrences to avoid giving too high a score to frequent LATs.

$F_k$, *failure score* of the LAT $k$, is the weighted error rate of $k$ for four, three, and two different models. $wrong(n,k)$ is the number of questions where the LAT is $k$ and those questions are wrong in $n$ different models. $tatal(k)$ is the total number of questions where the LAT is $k$. If the four models failed, for example, the question is considered difficult to solve for general MRQA models. On the other hand, if only one of the models could not solve it, it is regarded as an easy question to solve and it is not included in *failure score*. Because of this, we weight the error rate using $\frac{1}{2}^{(4-n)}$. If the four models fail, the weight would be 1, while if only two models fail, the weight would be 0.25.

### 3.2 Qualitative Analysis

We performed a qualitative analysis of XLNet with the same dataset to obtain some insights about why MRQA models cannot solve certain questions. We selected XLNet because XLNet is the current state-of-the-art on SQuAD. Another reason is that although some questions were not solved by XLNet but solved by other models, they did not provide additional insights. The failures made by XLNet were just caused by alias answers. We analyzed a random sample of 50 questions that were answered incorrectly by XLNet. The F1 score of those answers is lower than 0.1 to discard automatically the alias.

We analyzed the samples using two factors: the answer types and types of errors. First, to analyze the answer types, we categorized the questions into five groups: *clear LAT*, *abstract LAT*, *clear WH*, *abstract WH*, and *missing-answer-type*. These categories form part of a hierarchy as shown in Figure 1.

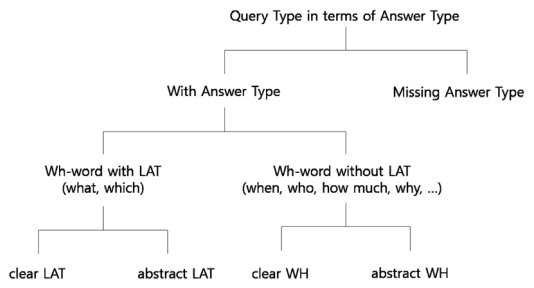


Fig. 1 Classification of questions

LAT is an explicit word that appears on *what-* and *which-* questions and specifies the type of answer needed. The group *wh-questions with LAT* can be divided into two sub-groups: *clear LAT* and *abstract LAT*. The former indicates a concrete object or concept like *country* and *school*, while the latter represents an abstract or overly broad concept like *goal*, *reason*, or *item*. We have the same subdivision for *wh-word without LAT* category. *Clear WH* questions are composed of *who*, *where, when*, or *how much/many*, which clearly represents the expected type of answer, and *abstract WH* questions are composed of *why* or *how* and therefore do not specify the type of answer as a noun-entity type. Second, we identified five types of error cases in the answers: alias problems, boundary problems, wrong contexts, wrong answer types, and modifiers. All these error cases are explained in Table 1.

## 4. Results

In section 4.1, a quantitative analysis is presented to show the general problems of MRQA models. In section 4.2, manual evaluation and qualitative analysis are shown to analyze the error cases in more detail. In the quantitative analysis, we analyzed two factors: wh-word and LAT.

We hypothesize that MRQA models leverage the answer type of the question to search entities in the text with the same type to select the answer. Therefore, if the answer type is ambiguous or too general, it is difficult for MRQA models to provide the right answer. This phenomenon occurs in both wh-type and LAT-type questions.

### 4.1 Quantitative Analysis with Squad Dataset

4.1.1 Difficulty of wh-types

In the first experiment shown in Figure 2, we see the proportion of questions per wh-word that can be solved with the four MRQA models we use. The analyzed questions words are: *when*, *who*, *how many*, *which*, *what*, *where*, *None*, and *why*.

According to the correct ratio of wh type, we can divide wh-type in four different groups.

*a) when, how many, who - easy to specify answer type*

The wh-words *when*, *how many*, and *who* show the highest correct ratio. The common characteristic

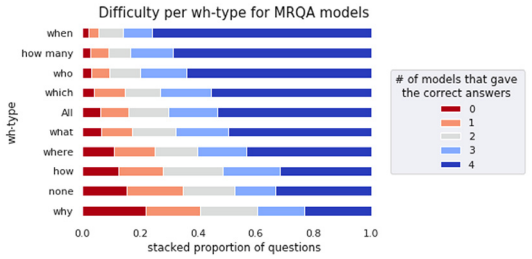


Fig. 2 Difficulty per wh-type for MRQA models

The graph shows which wh-type is more difficult for MRQA models. Wh-types are sorted increasingly by the level of difficulty. The legend indicates the number of models (BiDAF, DocumentQA, BERT, and XLNet) that gave the correct answers and the bar shows the proportion of each class in legend.

of these wh-words is that they can specify undoubtedly the expected answer type. For example, in the case of *when*, the answer must be something temporal like a date. *How-many*-questions require a numerical answer, and *who*-questions usually ask about people or organizations. Compared to other wh-types, these questions specify the expected answer type more concretely than the others, and thus, they are the easiest type of questions.

*b) what, which - depends on the LAT*

In the case of *what* and *which*, they show a similar correct ratio to the overall correct ratio (All). This is due to the fact that more than half of the questions are *what*-questions. In case of what- and *which*-questions, they usually include explicitly the answer type information inside the question. For example, in the question 'Which NFL team represented the AFC at Super Bowl 50?', the question explicitly states that the answer should be the name of a team. Even though *what*- and *which*-questions contains explicitly the LAT, the ratio of correct answers is lower than *when*, *who,* and *how many*. This is because there are many different LATs. Some of them are clear and some are abstract. In the case of the latter, MRQA models have problems as shown in section IV.A.2.

*c) where - granularity and ambiguity of location answer type*

The proportion of wrong *where*-questions is high even though the answer must be a location. However, the answer to the question 'where did you store the file?' could be 'in a pen drive', which is a location in this context, but it could not be a location in other contexts. This is because *location* can be a physical location like *country* or an abstract location, like in the example. Another reason for this phenomenon is the granularity of the answers. The answer provided by the model is correct but too general, i.e. a more specific answer can be given and thus, this more correct, and vice versa.

*d) None, why, how - requires a long answer*

The most difficult types of questions for MRQA models in terms of wh-type are *None*, *how*, and *why*. In the case of *None*-questions, since they do not have any wh-word in the question, the way of asking questions is a bit different. For example, 'Is corporal punishment increasing or declining in the South?'. Unlike asking an entity in the document that is suitable for the question, it asks to choose one of the options inside the question. This type of questions are less common in the dataset, and show a high failure rate because the models have to find the answer using a different strategy compared to other questions. The remaining type of questions, *how*, and *why* ask for an explanation. The high failure rate compared to other questions is due to three factors. Firstly, there is no link between the entity type of the answer and the answer type in the question. MRQA models can discover triggers to find answers inside the document for this type of questions, but it is difficult to generalize to all *how* and *why* questions. Secondly, since the answer to this type of questions is usually an explanation, they tend to be long, and thus, it is difficult to build an exhaustive enough golden answer list that covers all possible answers. Because of this, many answers are tagged as wrong because the predicted answer is not in the golden answer list even though it is correct. Thirdly, the predicted answers are not complete, i.e. they are only a partial answer. In conclusion, we think *None* questions are the most difficult because they require a deeper understanding of the evidence document to answer successfully.

4.1.2 Difficulty of LATs

As mentioned in section IV.B.2, *what*- and *which*-questions provide a precise granular answer type in the question. We define the Lexical Answer Type (LAT) as the first noun phrase after *what* and *which* words of the question. Using this simple heuristic, the LAT extraction accuracy is 93.68% in TriviaQA

according to a human evaluation. In dev set of SQuAD [3], there are total of 10,570 questions from which 6,735 are what- or which-questions. Among what- and which-questions, there are 2,231 unique LATs. The number of occurrences of LATs ranges from 1 to 286. To see a clear trend of the difficulty of each LAT type, the top 100 most frequent LATs are selected for the analysis.

In Figure 3, we sort the LAT by *difficulty score*, defined as the subtraction of the wrong ratio and the correct ratio.

$$Difficulty\ Score = \sum_{n=0}^{1} 2^{-n} \cdot (pr(n) - pr(4-n)) \quad (2)$$

where: $pr(x)$is the proportion of questions answered correctly by x models. We define that questions solved by less than half of the models are difficult while solved by more than half of the models are easy. We also give greater weight to extreme cases to emphasize the most difficult and easy questions.

*a) Clear vs. Abstract LAT*

The top 10 LATs that have the lowest difficulty score is *nationality, cost, building, year, school, date, city, forces, countries,* and *profession*. On the other hand, the top 10 LAT that has the highest difficulty score is *impact, reason, goal, article, inequality, disobedience, chloroplasts, role, dynasty,* and *way*. These two groups clearly show different characteristics. The former can specify an answer in a narrow range. For example, *nationality* can restrict the possible answers to just countries. We call *clear LAT* to these types of LAT. However, the latter LATs like *way* or *goal* are too general to specify the answer type by only using the LAT. We call *abstract LAT* to these types of LAT. First, we hypothesized that the high difficulty score of abstract LATs might be due to the lack of instances in the training set. However, as we will show in the sub-section *LAT Frequency*, there is no correlation between the frequency of a specific LAT in the training set and the accuracy in the dev set for that LAT. Therefore, we can conclude that MRQA models cannot handle properly queries with abstract LATs because of its semantics. One possible hypothesis of this phenomenon is that MRQA models learn how to use word embeddings to extract the semantics of the LAT to search candidate answers by entity type that corresponds to the LAT. If the question has a clear LAT, the word embedding of the LAT may contain information that can match it with the word embedding of the answer in the evidence document. However, if the question has an *abstract LAT*, it might be difficult that the word embedding of the answer has information that can match it with *the LAT.* This pattern widely appears in the dataset so it is plausible that the models learn it easily. In fact, humans also use this pattern to answer this type of questions. For example, if the LAT is *cat*, the answer candidate *Persian cat* may contain the information about cat in its embedding, but if the LAT is *thing*, the answer candidate *spider shooter* may not have information about the LAT inside its embedding because it is too general to keep.

*b) None LAT*

Several questions do not have LAT even though they are *what* or *which* questions. For example, 'A function problem is an example of what?'. According to Figure 3 (asterisk), *None* is one of the hardest LAT but not the hardest. Since *None* LAT does not provide information about the answer type, it should be the hardest type of question for MRQA models. However, unlike other *what-* and *which*-questions, these questions without LAT have a special pattern shown in Figure 4.

This finding shows that MRQA models are not only learning LAT patterns but also other patterns that can be easily found in the dataset. In most of the existing MRQA datasets, wh-word + (LAT) is a common pattern that can lead to the right answer and this is why MRQA models can base their answers on the answer type (wh-type or LAT).

1. The question ends with what → the noun or verb in front of what is a trigger to find the answer.
   ○ A function problem is an example of what?
2. The question ends with a preposition → usually the noun + preposition at the end of the question also appears in the document.
   ○ This network influenced later models of
3. The question ends with a verb → the verb at the end is the trigger to find the answer.
   ○ If someone is being taught at his place of residence, what is it called?

Fig. 4 Three patterns for None LAT questions

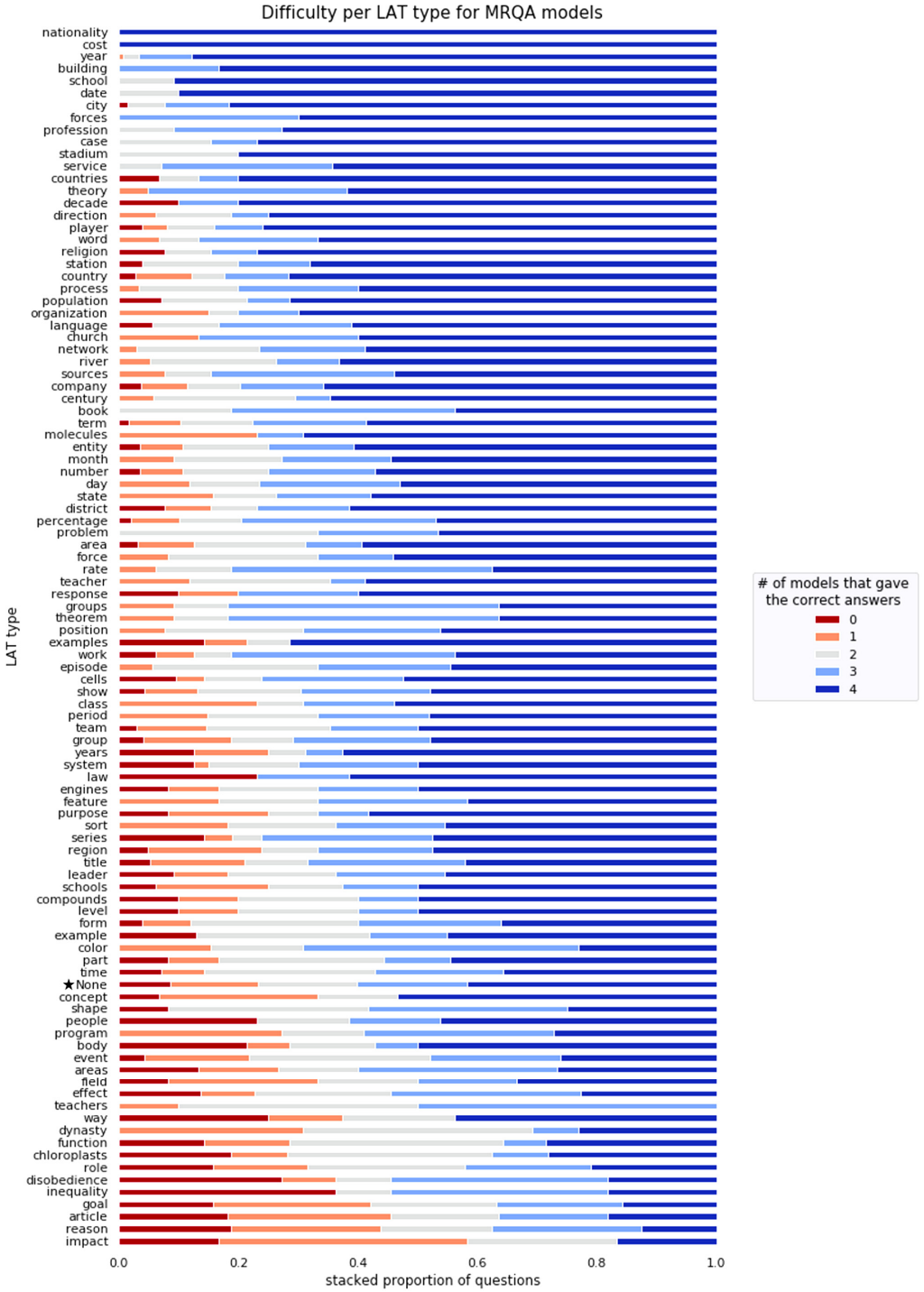


Fig. 3 Difficulty per LAT type for MRQA models

The graph shows which lat-type is more difficult for MRQA models. LAT types are sorted increasingly by the level of difficulty. The legend indicates the number of MRQA models (BiDAF, DocumentQA, BERT, and XLNet) that rendered the correct answers and the bar shows the proportion of each class in legend.

Table 2 Urgency of LAT

The frequent LATs in the dataset with a high failure rate should be corrected first to improve MRQA models. The table shows that non-lat is the hardest.

| | correct rate for lat in *n* different models | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *n* | 0 | 1 | 2 | 3 | 4 | | | | |
| LAT type | | | | | | freq | log_freq | wrong score | urgent |
| None | 0.114 | 0.143 | 0.191 | 0.157 | 0.394 | 690 | 6.537 | 0.234 | 161.5 |
| chloroplasts | 0.182 | 0.091 | 0.333 | 0.091 | 0.303 | 33 | 3.497 | 0.311 | 10.25 |
| company | 0.051 | 0.077 | 0.077 | 0.141 | 0.654 | 78 | 4.357 | 0.109 | 8.5 |
| country | 0.039 | 0.079 | 0.066 | 0.105 | 0.711 | 76 | 4.331 | 0.095 | 7.25 |
| goal | 0.211 | 0.211 | 0.211 | 0.211 | 0.158 | 19 | 2.944 | 0.368 | 7 |
| part | 0.086 | 0.086 | 0.286 | 0.114 | 0.429 | 35 | 3.555 | 0.2 | 7 |
| group | 0.042 | 0.146 | 0.104 | 0.229 | 0.479 | 48 | 3.871 | 0.141 | 6.75 |
| system | 0.122 | 0.024 | 0.122 | 0.22 | 0.512 | 41 | 3.714 | 0.165 | 6.75 |
| reason | 0.188 | 0.25 | 0.188 | 0.25 | 0.125 | 16 | 2.773 | 0.359 | 5.75 |
| role | 0.158 | 0.105 | 0.316 | 0.211 | 0.211 | 19 | 2.944 | 0.289 | 5.5 |
| way | 0.267 | 0.133 | 0.133 | 0 | 0.467 | 15 | 2.708 | 0.367 | 5.5 |
| effect | 0.13 | 0.087 | 0.217 | 0.348 | 0.217 | 23 | 3.135 | 0.228 | 5.25 |
| Doctor | 0.034 | 0.138 | 0.31 | 0.31 | 0.207 | 29 | 3.367 | 0.181 | 5.25 |
| example | 0.103 | 0 | 0.31 | 0.138 | 0.448 | 29 | 3.367 | 0.181 | 5.25 |
| impact | 0.167 | 0.333 | 0.333 | 0 | 0.167 | 12 | 2.485 | 0.417 | 5 |

4.1.3 LAT frequency

In this experiment, we check the existence of a correlation between the number of instances of each wh-word in the training set and its proportion of right answers in the dev set. As shown in Figure 5, there is no such correlation. The most common LATs, *year* and *country,* show similar behavior to infrequent LATs like *building,* with only 93 instances on the training set. All the questions with *building* as LAT can be successfully answered with at least one of the four models as shown in graph Figure 3. This experiment and the previous experiment show that the key to answering correctly a *what-* or *which-*question is the level of abstraction of the LAT rather than its number of instances on the training set since the word embedding of abstract LATs are more difficult to match with entities of the evidence document than clear LATs.

4.1.4 Urgency for Improving LAT Matching

In this experiment, we analyze the importance of LATs to improve the performance of the current MRQA models, i.e., how urgent we need to improve the performance of each LAT. To compute this *urgency* score, we used formula (2). As we can see in Table 2, it is crucial to improve the performance

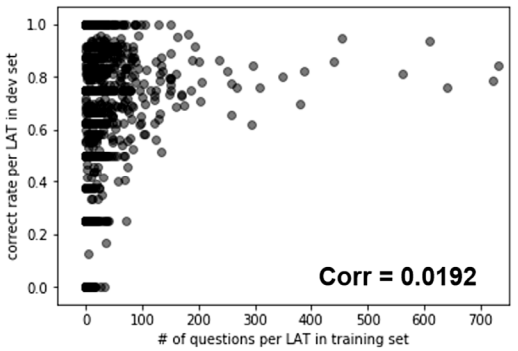


Fig. 5 Correlation between the number of instances per LAT in the training set and the correct rate per LAT in the dev set

Each dot indicates (lat, average accuracy) pair. The accuracy is calculated by averaging the accuracy values of the questions of a certain LAT.

of MRQA models on non-LAT questions. This type of questions are frequent and the current performance is low. On the other hand, *company*, and *country* are clear LATs, and thus, the proportion of questions that cannot be solved by any model is low. Nevertheless, since these LATs are extremely frequent, it is also important to improve the performance on them to improve the overall performance of

the models. Finally, as expected, the other LATs that are needed to improve are abstract. This result corroborates the experiment performed in Figure 3, current models have problems to answer questions with abstract LATs.

### 4.2 Qualitative Analysis with SQuAD Dataset

We first analyzed the type of questions that cannot be solved by BiDAF, DocumentQA, BERT, and XLNet. In the dev set of SQuAD, there are 422 questions out of 10,570 that cannot be solved by any of the four models (Figure 6). To see the characteristics of the questions that cannot be solved by MRQA models, we randomly sampled 20 out of these 422 questions. In Figure 7, unlike the previous analysis results from [9], the proportion of questions without AT is not that large compared to the questions with AT. We qualitatively analyzed the reason for this phenomenon and discovered that because of the addition of the new state-of-the-art model, XLNet, the overall accuracy increased, and thus, most of the errors are due to a dataset problem like non-exhaustive ground truth. We also found that in this sample there were no wrong answers due to a wrong AT, which suggests that our hypothesis, MRQA models base mainly their answers in the AT, is right. We analyzed the failure cases for each question type. In the case of clear LAT questions, a large part of failure cases is due to an alias and boundary problem. However, in relatively difficult questions like abstract LAT and questions without AT, there is a high proportion of errors due to a wrong context problem.

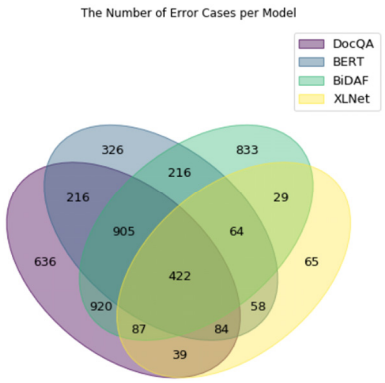


Fig. 6 Venn Diagram for the number of questions answered incorrectly by different subsets of the MRQA models

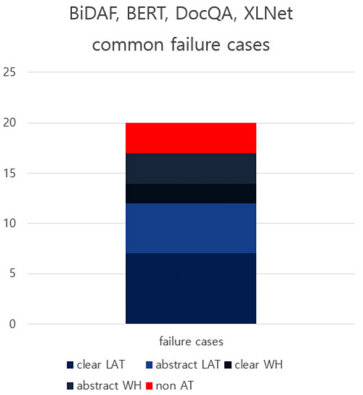


Fig. 7 Common Failure Cases for the four MRQA models The proportion of question-type is drawn on the graph. In this graph, non-AT does not show an impressive ratio of failure cases.

For further qualitative analysis, we use XLNet, as explained in *Methods* section (III-B). Most of the errors performed by XLNet in a random sample of questions are due to an alias problem. Therefore, qualitative analysis to determine the problems of MRQA models requires ignoring the alias problem. To do so, we apply a condition for sampling: the F1 score of the predicted answers must be below 0.1. In this way, we can avoid most of the alias problems.

#### 4.2.1 XLNet Failure Cases

In the analyzed random sample, we found out that most of the errors are due to the lack of understanding of the question or the evidence document by the model. For instance, the predicted answer to the question 'How do you pronounce Fresno?' is 'ash tree'. However, it is straight forward to find the answer in the evidence document. Some questions do require a deep understanding of the document and are even difficult for humans. For example, to answer the question 'What is the most critical resource measured to in assessing the determination of a Turing machine's ability to solve any given set of problems?' it is necessary to understand in detail the evidence document because the answer is not explicitly stated.

Another common type of error is due to using the LAT but not understanding the context surrounding the candidate answer. For example, the predicted answer for the question 'Which musical genre did the progressive folk-rock band Gryphon presented at a concert/lecture at the V&A?' is 'rock', but the golden

Table 3 XLNet Failure cases examples

| Error Type | Question<br>Context<br>(blue – right answer / red – prediction of model) | |
|---|---|---|
| | Right Answer | Prediction of Model |
| | Explanation for wrong reason | |
| Lack of understanding of the question | How do you pronounce Fresno? | |
| | Fresno (/ˈfrɛznoʊ/ FREZ-noh), the county seat of Fresno County, is a city in the U.S. state of California. … .The name Fresno means "ash tree" in Spanish, and an ash leaf is featured on the city's flag. | |
| | /ˈfrɛznoʊ/ FREZ-noh | ash tree |
| | The model answers the meaning of Fresno but not its pronunciation. | |
| Lack of understanding of the evidence document | What is the most critical resource measured to in assessing the determination of a Turing machine's ability to solve any given set of problems? | |
| | … decision problem A can be solved in time f(n) if there exists a Turing machine operating in time f(n) that solves the problem. Since complexity theory is interested in classifying problems based on their difficulty, one defines sets of problems based on some criteria. | |
| | time | difficulty |
| | The model needs to understand in detail the evidence document because the answer is not explicitly stated. | |
| Use of the LAT without understanding the surrounding context of the candidate answer | Which musical genre did the progressive folk-rock band Gryphon presented at a concert/lecture at the V&A? | |
| | … the V&A became the first museum in Britain to present a rock concert. The V&A presented a combined concert/lecture by British progressive folk-rock band Gryphon, who explored the lineage of mediaeval music and … | |
| | medieval music | rock |
| | The model is using the LAT to find the answer but is not understanding the document. | |
| Granularity of the answer | What does the template for bills passed by the Scottish Parliament include? | |
| | … Acts of the Scottish Parliament do not begin with a conventional enacting formula. Instead they begin with a phrase that reads: "The Bill for this Act of the Scottish Parliament was passed by the Parliament on [Date] and received royal assent on [Date]" | |
| | The Bill for this Act of the Scottish Parliament was passed by the Parliament on [Date] and received royal assent on [Date] | a phrase |
| | The predicted answer is correct but does not give the expected information. | |

answer is 'medieval music'. In this example, it is easy to see that the model is using the LAT to find the answer but is not understanding the document. Among the 50 randomly selected questions, 26 questions are wrong because of this. Another common error we found is due to the granularity of the answer. For example, the predicted answer for the question 'What does the template for bills passed by the Scottish Parliament include?' is 'a phrase', which is right but at the same time, it does not give much information. The golden answer is actually 'The Bill for this Act of the Scottish Parliament was passed by the Parliament on [Date] and received royal assent on [Date]'. This answer has more information and thus, it is more complete and accurate. We found this type of error in five questions. The remaining questions were right, but the predicted answer was not in the list of golden answers or the golden answer was wrong due to a problem in the dataset. All cases are described in Table 3.

## 5. Discussions and Conclusions

We discussed some of the limitations of recent Machine Reading Question Answering (MRQA) models. First, current MRQA models do not understand the questions and evidence documents but exploit some easy patterns that occur in the datasets. The most common pattern is the matching between the answer type (AT) of the question and the entity type of the answer. This implies that the models only learn how to detect shallow patterns. To advance in the reading comprehension capabilities of these models, we require new better architectures and datasets. Second, since the models heavily rely on AT pattern, when the AT is too abstract or general, the models struggle to find the right answer. We think there are two ways to overcome this problem. One is to create better word embeddings for matching answer types. The other option is to increase the number of instances of abstract LATs in the training set to help the models to capture the pattern of abstract LATs.

Our main contribution is an analysis of the types of questions in SQuAD [3] that are difficult to answer by MRQA models. We have shown through quantitative and qualitative experiments that questions

with an abstract LAT and abstract wh-word questions and those without a wh-word or LAT are the most difficult to answer for MRQA models. This also implies that these models are using the AT to find the answers in the evidence document. We also showed that SQuAD is an easy dataset to solve because most of the questions do contain an AT, so MRQA models can use a simple pattern to obtain a high performance even though they do not understand the evidence documents.

One of the limitations of our work is that we focused only on SQuAD and did not analyze other new datasets like HotpotQA [7], and Natural Questions [8]. Because of this, we cannot generalize the flaws of SQuAD to other datasets. We plan to analyze other datasets to generalize the limitations of MRQA datasets as a prior step to build a new dataset. This new dataset will focus on questions without AT. We believe that this type of questions can contribute to the creation of new models that may really understand documents, instead of using simple patterns like AT.

## References

[1] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 4171-4186, 2019

[2] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. and Le, Q. V, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *arXiv preprint arXiv:1906.08237*. 2019.

[3] Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383-2392, 2016.

[4] Weissenborn, D., Wiese, G., & Seiffe, L. Making neural qa as simple as possible but not simpler, *arXiv preprint arXiv:1703.04816*, 2017.

[5] Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H, "Bidirectional attention flow for machine comprehension," International Conference on Learning Representations, 2017.

[6] Clark, Christopher, and Matt Gardner, "Simple and Effective Multi-Paragraph Reading Comprehension," *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 845-855, 2018.

[7] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W, "Hotpotqa: A dataset for diverse, explainable multi-hop question answering," *Conference on Empirical Methods in Natural Language Processing*, Vol. 1, pp. 2369-2380, 2018.

[8] Kwiatkowski, Tom, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein et al., "Natural questions: a benchmark for question answering research," *Transactions of the Association for Computational Linguistics 7*: 453-466, 2019.

[9] Puerto San Roman, Haritz and Lim, Doyeon, "Analysis Answer Type Application Ability of State-of-the-Art Reading Comprehension Models for Question Answering Task," *Korea Computer Congress*, 2019.

[10] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD," *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 2: Short Papers)*, pp. 784-789, 2018.

[11] Joshi, Mandar, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer, "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension," *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601-1611, 2017.

**Doyeon Lim** is a master's student in the School of Computing at Korea Advanced Institute of Science and Technology (KAIST) and research assistant in IR&NLP Lab at KAIST. She graduated from the Ulsan National Institute of Science and Technology with a degree in Biomedical Science and Electrical Engineering in August 2019. lays in the application of Machine Learning to Natural Language Processing, specifically, Question Answering, Personal Recommendation.

**Haritz Puerto San Roman** is a master's student in the School of Computing at Korea Advanced Institute of Science and Technology (KAIST) and research assistant in IR&NLP Lab at KAIST. He graduated from the University of Malaga with a degree in Computer Science in July 2017. His research lays in the application of Machine Learning to Natural Language Processing, specifically, Question Answering, Question Generation, and Open Information Extraction.

**Dr. Sung-Hyon Myaeng** is a professor of Computer Science in the School of Computing and the head of Web Science & Technology Division at Korea Advanced Institute of Science and Technology (KAIST). He is also the Director of KAIST-Microsoft Research Collaboration Center (KMCC). Previously he was on the faculty at Syracuse University, USA, where he was granted tenure in 1994. He earned his MS and Ph. D. from Southern Methodist University, Texas, USA in 1985 and 1987, respectively. His research has been in the intersection between lexical & semantic aspects in natural language processing and unconventional search techniques in information retrieval.