IMPORTANCE OF THE SINGLE-SPAN TASK FORMULATION TO EXTRACTIVE QUESTION-ANSWERING

Marie-Anne Xu[1] and Rahul Khanna[2]

[1]Crystal Springs Uplands School, CA, USA
[2]University of Southern California, CA, USA

ABSTRACT

Recent progress in machine reading comprehension and question-answering has allowed machines to reach and even surpass human question-answering. However, the majority of these questions have only one answer, and more substantial testing on questions with multiple answers, or multi-span questions, has not yet been applied. Thus, we introduce a newly compiled dataset consisting of questions with multiple answers that originate from previously existing datasets. In addition, we run BERT-based models pre-trained for question-answering on our constructed dataset to evaluate their reading comprehension abilities. Among the three of BERT-based models we ran, RoBERTa exhibits the highest consistent performance, regardless of size. We find that all our models perform similarly on this new, multi-span dataset (21.492% F1) compared to the single-span source datasets (~33.36% F1). While the models tested on the source datasets were slightly fine-tuned, performance is similar enough to judge that task formulation does not drastically affect question-answering abilities. Our evaluations indicate that these models are indeed capable of adjusting to answer questions that require multiple answers. We hope that our findings will assist future development in question-answering and improve existing question-answering products and methods.

KEYWORDS

Natural Language Processing, Question Answering, Machine Reading Comprehension

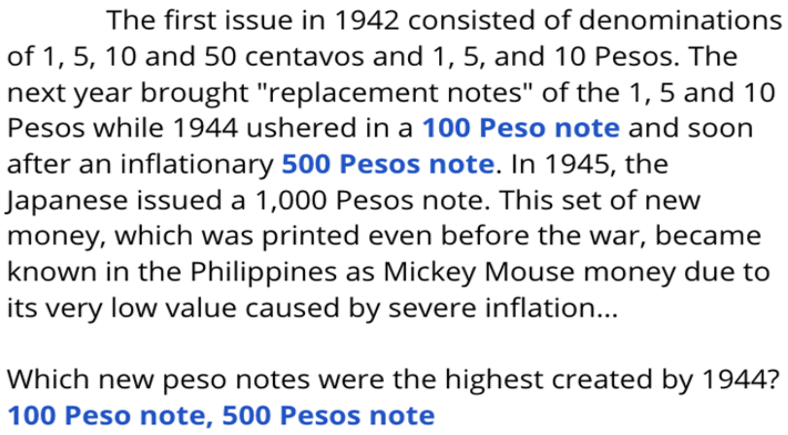
1. INTRODUCTION

Machine Reading Comprehension (MRC), particularly extractive close-domain question-answering, is a prominent field in Natural Language Processing (NLP). Given a question and a passage or set of passages, a machine must be able to extract the appropriate answer or even set of answers from the passage(s). Solving this task has various real-world implications, particularly in industry areas such as customer support. Some application examples include chatbots, voice assistants, and automated customer service. Using these applications can greatly increase efficiency for both companies and customers by reducing time spent hunting for answers that a machine can find in seconds.

Many groundbreaking question-answering datasets such as the Stanford Question Answering Dataset, SQuAD [1], consist of only single-span question-answer pairs, or answers that require only one extraction. Numerous datasets have been created with several answer categories, NewsQA [2], DuReader [3], MS MARCO [4], DROP [5], however, the majority of the answers are single-span. Models that are trained on these datasets are therefore primarily focused on extracting just a single answer, possibly precluding their effectiveness when multiple answers are

David C. Wyld et al. (Eds): CSEA, DMDBS, NSEC, NETWORKS, Fuzzy, NATL, SIGEM - 2020
pp. 107-116, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101809

required. For example, one sample of a multi-span question-answer pair along with its relevant passage from the DROP dataset is shown in Figure 1.



The first issue in 1942 consisted of denominations of 1, 5, 10 and 50 centavos and 1, 5, and 10 Pesos. The next year brought "replacement notes" of the 1, 5 and 10 Pesos while 1944 ushered in a **100 Peso note** and soon after an inflationary **500 Pesos note**. In 1945, the Japanese issued a 1,000 Pesos note. This set of new money, which was printed even before the war, became known in the Philippines as Mickey Mouse money due to its very low value caused by severe inflation...

Which new peso notes were the highest created by 1944?
100 Peso note, 500 Pesos note

Figure 1: A shortened DROP dataset passage with its multi-span question-answer pair.

Various models such as Retro-Reader [6], SA-Net, and ALBERT [7] have already surpassed human performance on MRC datasets like SQuAD 2.0 [8], which introduces questions that may not have answers given the passage(s). However, model performance on datasets such as Discrete Reasoning Over Paragraphs (DROP) have not yet matched human quality, and models trained on datasets like DuReader are even further away from reaching human performance. These contrasting results show that it is unclear whether or not models are able to execute machine reading comprehension. More recently, models such as the Multi-Type Multi-Span Network [9] are specially designed to extract either one or multiple answer spans. While a specialized model is capable of returning multi-span answers, we seek to investigate if the current state-of-the-art models can adapt without fine-tuning to produce similar results.

Thus, this research project proposes to assess the performance of the current state-of-the-art models when evaluated on only the multi-span questions of existing datasets. By exploring the MRC abilities of models trained on single-span extraction, we can determine if the model is simply concentrating on returning only one answer or if it is actually processing and comprehending the task of question-answering. Future researchers will be able to use this work in order to identify where question-answering models can be improved and recognize the limitations of using a model trained on a single-span dataset. Additionally, single-span answers could potentially be overly specific for a given question, thus exploring multi-span answer question-answering can potentially provide the end-user with more information to answer their questions. The new multi-span dataset compiled from the DROP and NewsQA datasets is also available for future research.

2. RELATED WORKS

As previously mentioned, the majority of existing question-answering pairs have a single answer. The type of dataset we are particularly interested in for our evaluation is extractive closed-domain question-answering. The appropriate answer(s) must be directly extracted from only the given passage(s). One example of this kind of dataset is the SQuAD dataset, with over 100,000 single-span questions, making it larger than most previous question-answering datasets. SQuAD has several answer types, such as Date, Person, and Location, and its passages cover an extensive number of subjects. SQuAD 2.0, the latest version of SQuAD, combines the original 100,000 questions with new unanswerable questions, forcing models to learn when to abstain from answering. Similarly, the DROP and NewsQA datasets are also extractive, closed-domain datasets that have a small percentage of multi-span question-answer pairs (6.0% for DROP and 5.68% for NewsQA). The DROP dataset contains questions that may need numerical operations (e.g. “Who threw the longest touchdown pass?”) or entity co-referencing (tracking the multiple appearances of an entity). Compared to SQuAD, DROP has more complex passages and questions that require more complicated reasoning. Like SQuAD 2.0, NewsQA has more ambiguous questions and questions without answers, and its answer types mirror those of SQuAD (Numeric, Clause Phrase, Date/Time, Location, etc). However, the human performance of NewsQA is much lower than the previous two datasets, despite it being a similar size. While other datasets such as the MS MARCO and DuReader datasets have multi-span answers, they are either generative (answers cannot be directly extracted), or are in another language like Chinese.

Some of the most popular NLP models are BERT [10] and its variations, RoBERTa [11] and ALBERT [7]. BERT, which stands for Bidirectional Encoder Representations from Transformers, is capable of executing a variety of general NLP tasks, such as next sentence prediction. The key feature of BERT is that it can look at the context on both sides of a given text span, hence the part “bidirectional”. RoBERTa, or Robustly optimized BERT approach, introduces alternate strategies for the BERT training process in order to improve performance. ALBERT (A Lite BERT) contains fewer parameters than BERT, reducing training time and memory restrictions while simultaneously maintaining and even producing better results.

3. Purpose

Multi-span question-answering can be used to improve existing applications of question-answering and NLP in general. As mentioned earlier, MRC plays a key role in industry, forming the foundation for tools such as virtual assistants. However, these machines are still flawed. If a certain question posed by a user requires multiple answers, the machine needs to be able to find different, quality answers in order to be as productive and helpful as possible. Models that are trained on mainly single-span questions may not exhibit high-quality performance when faced with multi-span questions. Using our evaluations, future research will improve question-answering models and hopefully implement them to refine existing applications. If the models are capable of answering multi-span questions posed by an end-user, they will minimize the user’s time spent looking for an answer that the machine was not able to return. By contributing to the field of MRC with our evaluations, we also hope to further the application of MRC and NLP as a whole in the real world.

4. Methods

To analyze model performance on multi-span question answering, we collect all question-passage-answer triplets that have multi-span answers from the DROP and NewsQA datasets. The newly compiled, strictly multi-span dataset consists of almost 30K questions. As shown in Table 1, most passages are quite long, while answers are only a few words.

Table 1. Multi-span dataset statistics.

Statistic	
Number of Questions	29288
Avg question len (words)	7.10
Avg passage len (words)	546.54
Avg answer len (words)	4.69
Avg answers/question	2.22

Refer to Figure 2 for some details on the top first and second words of the questions. Similar to a pie chart, the first words are represented by the inner circle, and the second words are shown as subcategories in the outer ring. 16,674 of the total 29,288 questions are represented here, roughly 57% of the entire dataset. The first words are dominated by "what," "who," and "where" questions; the "what" section forms approximately 59% of these 16,674 questions, "who" makes up 14%, "where" 10%, and "which" 8%. The unusual "players" section under "which" most likely originates from the large number of National Football League (NFL) questions in the DROP dataset. For example, "which players scored touchdowns longer than 20 yards?" would fall under the first word category "which" and second word category "players".

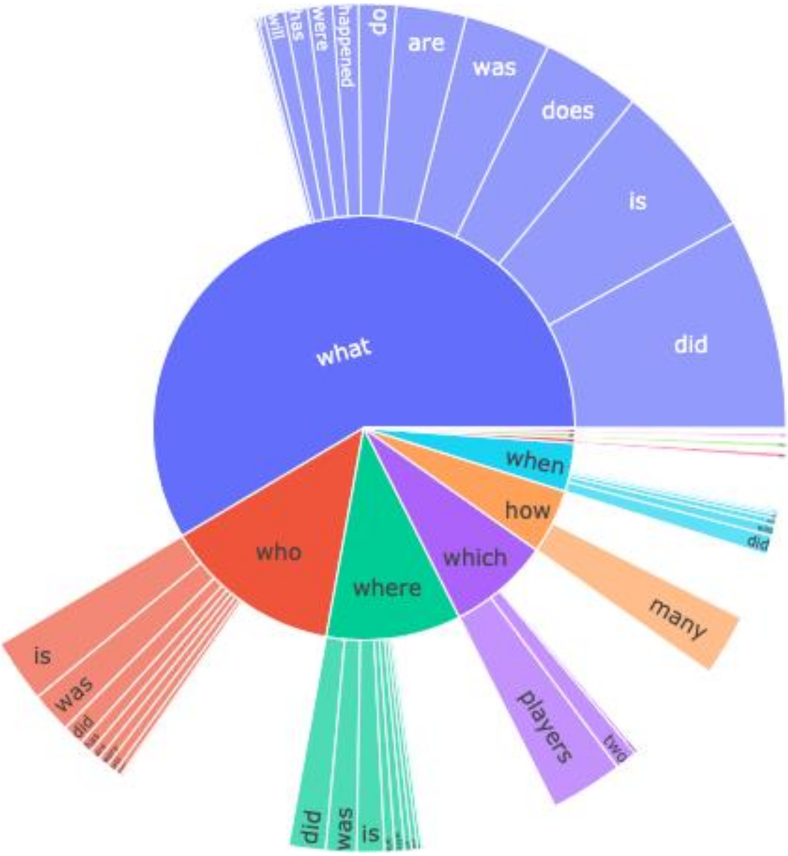


Figure 2: The top 10 first and top 14 second question words of the multi-span dataset.

The models run on this dataset are the most downloaded BERT-base [14], BERT-large [15], RoBERTa-base [16], RoBERTa-large [17], ALBERT-base [18], and ALBERT-xlarge [19] models fine-tuned on the SQuAD2.0 dataset. We choose these models because they have already been fine-tuned for single-span extractive question-answering and can be easily modified and evaluated on their multi-span question-answering abilities. We follow the standard preparation of question and passage pairs for the extractive task, [CLS] question text [SEP] passage text. As BERT based models have a token limit of 512 tokens, we follow common practice of truncating all constructed sequences to the 512 token limit, this affects 19,837 of our question, passage pairs. Due to limited runtime, we run all models on the 9,451 shorter question, passage pairs and the three base models on the entire dataset. Evaluation runtime is approximately three days using a Google Cloud Platform Compute Engine CPU. After passing the tokens into the model, the model produces the start and end scores for each token; for example, the token with the highest starting score marks the start of the best span, and the token with the highest end score is the last token in the span. We utilize a relatively naive, greedy approach by extracting the top n nonoverlapping answer spans with the best scores. The number of predicted answer spans is determined by the number of true answer spans. We construct the predicted answers by concatenating the relevant tokens; all of the tokens between the start and end token are part of the predicted answer span. If the model is unable to find an acceptable span, it returns an empty string.

To assess the models, we create two arrays that are the same length as the passage's character length. One array is for the predicted answers while the other is for the true answers. For each answer span, we find its location in the passage by string matching; the answer span is a substring of the passage. In order to accurately match the strings, we must lowercase the original passage as BERT and ALBERT answer spans are not case-sensitive. We do not lowercase for RoBERTa because the most downloaded models, and subsequently the ones we used in this experiment, are case-sensitive. Characters in the passage that are part of an answer are labeled as a one (referred to as Class 1) while the rest are zeroes (Class 0). The result is an array of ones and zeros that is the same length as the character length of the passage. We can then compare the indices of the characters in the true and predicted answers, or the ones in the arrays. Due to a time constraint, we choose the first occurrence of the answer in the passage if it appears multiple times. This does not affect the evaluations because the answer spans are neither overlapping nor repeated.

For each predicted answer set, we calculate the average exact match: how many predicted answers are exactly the same as the true answers? As stated above, we lowercase the true answers for BERT and ALBERT before matching and do not lowercase for RoBERTa. We also calculate the micro and Class 1 precision, recall, and F1 scores between the true and predicted binary arrays that we created earlier. Precision measures the number of characters in the predicted answers that are also true; micro-precision calculates the global precision while Class-1 reports the precision for each answer. Conversely, recall measures the number of characters in the true answers that are also in the predicted answers. We report the precision, recall, and F1 scores because they judge how similar our predicted answers are to the true answers when they are not identical.

Our final metric is BLEU [12], an automatic evaluation method that originated from the field of machine translation. BLEU compares the number of *n-grams*, or set of words, present between the candidate (predicted answer, in our case) and reference (true answers) sentences. BLEU also penalizes the candidate (predicted) sentence when it is shorter than the reference(s), called the brevity penalty. We use the BLEU-1 score—the majority of answers are only one word, thus we only use unigrams when computing our BLEU scores. We calculate the BLEU scores between the predicted and true answer spans for each question and then find the arithmetic mean. BLEU differs from the previous metrics, such as precision, in that it introduces the brevity penalty and also measures by word, while precision measures by character. We can then compare these all of the mentioned metrics to the original DROP and NewsQA evaluations to determine the performance of our models.

5. RESULTS AND DISCUSSION

Of the three types of models, RoBERTa stands out as consistently exhibiting strong scores in all evaluation metrics. Other models like BERT-base and ALBERT-xlarge had peaks in certain metrics, like exact match and F1 respectively, but their remaining metrics were not as prominent. RoBERTa-large had the more consistently high scores, most notably the highest Class-1 F1 and recall scores. Class-1 metrics are calculated when we only consider the true (Class-1) characters in our binary arrays. A few key metrics (EM, micro-F1, and BLEU) are displayed in Figure 3 and Table 2. The remaining evaluations (precision, recall) for each model can be seen in Table 3.

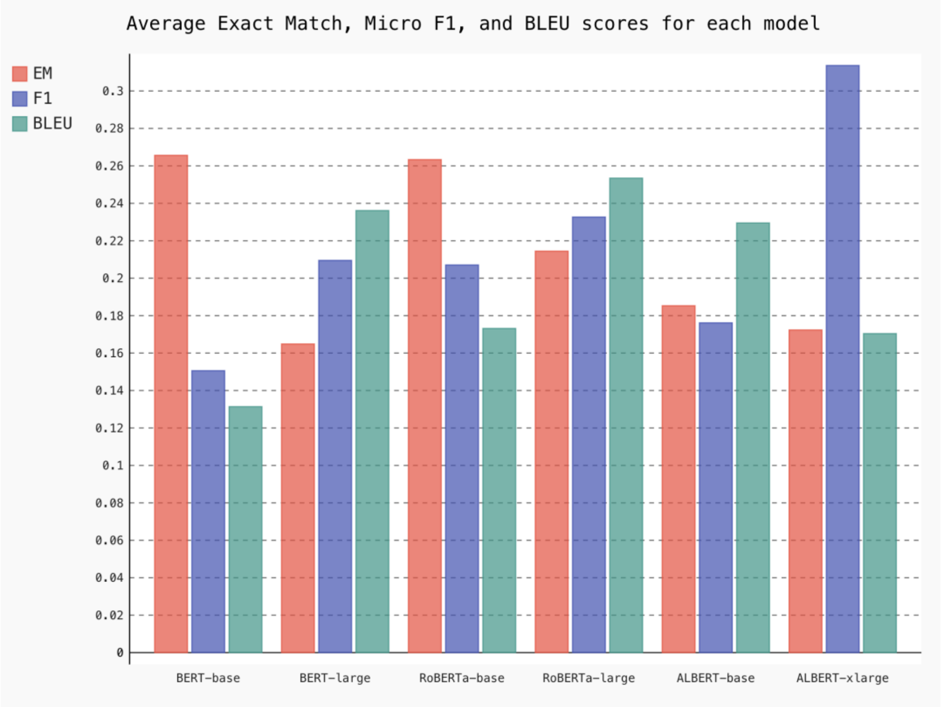


Figure 3: EM, micro-F1, and BLEU-1 scores for each model. The large models are run on a smaller version of the dataset (32%); base models are run on the entire dataset.

Table 2: A table representation of Figure 3.

	EM	**F1**	**BLEU-1**
BERT-base	26.561	15.057	13.138
BERT-large	16.484	20.949	23.612
RoBERTa-base	26.333	20.707	17.314
RoBERTa-large	21.445	23.267	25.342
ALBERT-base	18.527	17.613	22.949
ALBERT-xlarge	17.233	31.360	17.039
Average	21.097	21.492	19.899

Comparing our metrics in Figure 3 and Table 2 to the overall BERT scores of the entire DROP dev dataset, we can see that the evaluations of our six models are worse. This is expected, as the majority of the DROP dataset consists of single-span answers. The BERT scores for SQuAD-style Reading Comprehension on the DROP dev dataset are 30.10 (EM) and 33.36 (F1), using the BERT-large model. Additionally, the BERT model used in the DROP dataset was slightly fine-tuned while our BERT models were not.

When we observe the evaluations of the original multi-span subset of DROP, we find that the BERT scores listed in their paper are 0 (EM) and 25.0 (F1). In comparison, our BERT scores are

26.561 (EM) and 15.057 (Micro F1). Our exact match score for BERT-large, and all six models in general, are much higher, while our F1 scores are similar. We reason that this result holds as the multi-span DROP questions only made up around 20% of our dataset while the other 80% came from NewsQA, and EM scores on the NewsQA dataset are much higher. However, BERT-based models were not run on the original NewsQA dataset, so we cannot do a like-for-like evaluation. Our greedy strategy of selecting the multiple top answers has most likely also affected the performance scores of our models.

While NewsQA was not run on a BERT-based model, the alternate models stated in the original paper, mLSTM and BARB, produced EM scores of 34.4% and 36.1% respectively. The F1 scores for both models were 49.6%. Again, the majority of NewsQA consists of multi-span question-answer pairs, so our lower scores are expected. While we cannot definitively conclude the performance of these BERT-based models aren't affected by the change of task, we can see that the models are adapting to multi-span question-answering to a high degree, as the EM and F1 scores are not extremely low.

Table 3: Class-1 metrics, which focus on the true (Class-1) characters rather than words.

	Class-1 F1	**Class-1 Precision**	**Class-1 Recall**
BERT-base	18.418	24.892	20.677
BERT-large	28.588	34.509	35.028
RoBERTa-base	24.766	33.690	27.028
RoBERTa-large	34.004	42.586	43.010
ALBERT-base	27.492	31.539	39.406
ALBERT-xlarge	26.199	33.861	28.274
Average	26.578	33.513	32.237

In Table 3, the class-1 metrics are calculated using the previously mentioned methods. Some notable metrics include the high precision and recall of RoBERTA-large, 42.586 and 43.010 respectively, as well as the high recall of ALBERT-base (39.406). The overall recall of the models is the highest of the general metrics, and we judge recall to be the most important metric because it measures the number of true characters that are in the predicted answers. Although the predicted answers may not be exactly the same as the true ones (which the exact match score penalizes), recall checks for the presence of the true characters. Framing our evaluations as a sequence tagging task, we look at the performance of our models as sequence taggers, examining in particular their performance when an answer is the true label for a character (i.e. the character is within the answer). In this regard, class-1 recall is a key statistic, as this would show that the model is able recover the true answers' character tokens, while allowing it to potentially start the span a bit earlier or later than the annotated answer, which does not necessarily affect our recall negatively. Because we are checking for the presence of the true characters and judging the models' abilities to identify them, we do not focus much on the additional characters in the span.

We expect RoBERTa to perform better than BERT, as it is pre-trained for a longer period of time. What is unexpected, however, is that ALBERT does not surpass RoBERTa's performance. There are several potential reasons for this, one being that ALBERT is much smaller in size than RoBERTa, but we leave this exploration to future research.

Generally, the larger models return higher scores than their base counterparts. This is expected, as the larger models usually outperform the base versions, even in non-question-answering tasks. However, one exception to this trend is that the exact match scores of the large models are lower than the bases. Another notable comparison is that despite ALBERT-xlarge being several times larger than ALBERT-base, various ALBERT-xlarge metrics are either close to or lower than ALBERT-base's, like the exact match, class-1 F1, precision, and recall, micro-recall, and BLEU-1 scores. The remaining metrics, the micro-F1 and micro-precision scores, are much higher and match our expected trend that increased size implies improved performance.

Our overall exact match and F1 scores, especially compared to the DROP and NewsQA scores, reveal that the six models are capable of adapting and returning more than one answer span. Because our models can produce scores that are not significantly lower than previous scores and are even better in some cases, we can infer that the models are indeed adjusting to multi-span question-answer pairs.

6. CONCLUSIONS

Because the majority of existing machine question-answering datasets consist of single-span question-answer pairs, we seek to evaluate state-of-the-art model performance on multi-span questions. We have assembled a new multi-span question-answer dataset from the DROP and NewsQA datasets. We have observed the performance of six BERT-based models pre-trained for single-span question-answering when run on this compiled multi-span dataset. We find that RoBERTa has consistently higher scores than BERT and ALBERT, perhaps due to its alternate training strategies, and we also note that the larger variations of each model perform better than the base counterparts, although at the cost of increased runtime. When comparing the EM and F1 scores of our BERT-based models to the BERT-scores of the parent DROP dataset, we find that the EM scores have improved significantly and the F1 scores are similar, although slightly lower. Based on the unbalanced distribution of DROP and NewsQA questions, we also look at the scores of the other parent dataset, NewsQA; although not from BERT, we see that our scores are not drastically lower. Because our BERT models have not been fine-tuned to our multi-span dataset while the BERT model for DROP evaluation has, this difference in scores still allows us to conclude that task formulation does not drastically reduce model performance. We discover that current state-of-the-art models are capable of adapting to multi-span extractive question-answering and are not structurally limited to single-span extraction. We hope that with this information, future research can implement multi-span question-answering into real-world applications to improve efficiency in industry and daily life.

7. FUTURE WORK

Some potential future research directions include fine-tuning a model to answer multi-span questions and conforming to more standard evaluation metrics (such as CIDER [13]) We also suggest exploring alternate, more robust extraction methods that are better than our naive greedy approach. Another prospective project involves more detailed performance evaluation on certain subsets of this multi-span dataset, such as the "who," "what," "when," "where," "which," and "how" question subcategories. Additionally, more work can be done in lengthening the extracted spans in order to use the standard BLEU-4 evaluation method. In hopes of furthering MRC and NLP, we also leave this multi-span dataset available for future research.

ACKNOWLEDGEMENTS

We would like to thank the Summer Stem Institute for providing the opportunity to conduct this research.

REFERENCES

[1] 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, Austin, Texas, November 2016. Association for Computational Linguistics.

[2] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. "NewsQA: A machine comprehension dataset". In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[3] Wei He, Kai Liu, Jing Liu, YajuanLyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. "DuReader: a Chinese machine reading comprehension dataset from real-world applications". In P*roceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[4] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. "MS MARCO: A human generated machine reading comprehension dataset". *CoRR*, abs/1611.09268, 2016.

[5] DheeruDua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs". In *Proceedings. of NAACL*, 2019.

[6] Zhuosheng Zhang, Junjie Yang, and Hai Zhao. "Retrospective reader for machine reading comprehension", 2020.

[7] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "Albert: A litebert for self-supervised learning of language representations". In *International Conference on Learning Representations*, 2020.

[8] Pranav Rajpurkar, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD". In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[9] Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. "A multi-type multi-span network for reading comprehension that requires discrete reasoning". 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 1596–1606, Hong Kong, China, November 2019. Association for Computational Linguistics.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: pre-training of deep bidirectional transformers for language understanding". *CoRR*, abs/1810.04805, 2018.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and VeselinStoyanov. "Roberta: A robustly optimized BERT pretraining approach". *CoRR*, abs/1907.11692, 2019.

[12] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation". In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[13] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation". *CoRR*, abs/1411.5726, 2014.

[14] BERT-base: https://huggingface.co/twmkn9/bert-base-uncased-squad2

[15] BERT-large: https://huggingface.co/deepset/bert-large-uncased-whole-word-masking-squad2

[16] RoBERTa-base: https://huggingface.co/deepset/roberta-base-squad2

[17] RoBERTa-large: https://huggingface.co/ahotrod/roberta_large_squad2

[18] ALBERT-base: https://huggingface.co/twmkn9/albert-base-v2-squad2

[19] ALBERT-xlarge: https://huggingface.co/ktrapeznikov/albert-xlarge-v2-squad-v2

[20] Jiahua Liu, Wan Wei, Maosong Sun, Hao Chen, Yantao Du, and DekangLin. “A multi-answer multi-task framework for real-world machine reading comprehension”. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2118, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[21] LluísMàrquez, Pere Comas, Jesús Giménez, and NeusCatalà. “Semantic role labeling as sequential tagging”. In *Proceedings of the Ninth Confer-ence on Computational Natural Language Learning (CoNLL-2005)*, pages 193–196, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[22] Pradeep Dasigi, Nelson F. Liu, Ana Marasovíc, Noah A. Smith, and Matt Gardner. “Quoref: A reading comprehension dataset with questions requiring coreferential reasoning”, 2019.

[23] Subeh Chowdhury, Michael O’Sullivan. “A fuzzy logic-genetic algorithm approach to modelling public transport users’ risk-taking behaviour”. *Transportation Planning and Technology 41:2*, pages 170-185.

[24] G. Megali, D. Pellicano, M. Cacciola, S. Calcagno, M. Versaci, and F. C. Morabito, "Ec Modelling and Enhancement Signals in Cfrp Inspection," *Progress In Electromagnetics Research M*, Vol. 14, 45-60, 2010.

AUTHORS

Marie-Anne Xu is a student with a keen interest in computational linguistics. She has competed in the North American Computational Linguistics Open Competition twice and qualified for the invitational round in 2020. In her spare time, she has participated in coding events such as hackathons and Google Code-In. She has received the National Center for Women & Information Technology Award for Aspirations in Computing National Certificate of Distinction.

Rahul Khanna is a Masters Student at USC's Viterbi School of Engineering and a Researcher at USC's Intelligence and Knowledge Discovery Research Lab. He is primarily interested in semi-supervised, self-supervised and human-in-the-loop learning for Information Extraction, but also explores the areas of Common Sense Reasoning and Question-Answering. Prior to USC, Rahul earned his B.S. from Columbia's Fu School of Engineering and Applied Sciences and spent three years working as machine learning engineer/datascientist for Vimeo and Impact.