FLUD: Expert-curated large-scale machine comprehension dataset with advanced reasoning strategies

Yi-Chih Huang[1,*], Yu-Lun Hsieh[1], Yi-Yu Lin[1], Teo Lin Hui[1], Hung-Ying Chu[1], Wen-Lian Hsu[1]

[1] Science & Technology Policy Research and Information Center, National Applied Research Laboratories, Taipei 10636, Taiwan
[*] Correspondence: yichuang@narlabs.org.tw

Abstract

We introduce the Formosa Language Understanding Dataset (FLUD), a large open-domain Traditional Chinese machine reading comprehension dataset curated by professionals. FLUD contains more than 15,000 textual question-answer pairs and corresponding articles from Wikipedia, news, and elementary school textbooks. The questions are in the form of multichoice or short-answer problems, and the level of difficulty is evaluated by organizers of official language proficiency tests. In addition, we incorporate task-oriented multiround dialogue and record parts of the question-answer pairs spoken by humans to extend the breadth of this dataset. The aim is to design a more challenging dataset that requires advanced reasoning beyond straightforward span extraction techniques to answer the questions correctly. FLUD was used in two public machine-learning competitions, in which we conducted human evaluations on the difficulty of this corpus. Human evaluations on the accuracy of multichoice and short-answer problems amount to 89.6% and 62.7%, whereas the best machine performances are 53.7% and 40.8%, respectively. The public competitions energize the scientific and engineering community and the public to develop a sense of the possibilities and an urgent commitment to accelerate progress.

1. Introduction

Language understanding technology is receiving considerable attention from the artificial intelligence (AI) and natural language processing (NLP) community and people in the general public. The application ranges from text classification to question answering (QA) and multiturn conversations. Furthermore, the machine needs to categorize documents, produce casual conversations, and respond to user questions regarding specific topics or tasks to be comprehensive. In order to boost development in this direction, we construct the Formosa Language Understanding Dataset (FLUD), of which the level of difficulty is set to match "Band C" of the "Test of Chinese as a Foreign Language" (TOCFL) [1]. This level roughly corresponds to "Advanced" in other language proficiency tests such as the Common European Framework of Reference for Languages. The corpus contains three different tasks: around 15,000 multiple-choice quizzes, more than 700 short-answer questions, and 20 task-oriented multiturn dialogues. The majority of this dataset contains audio recordings as well as accompanying textual content. Moreover, we design the questions with special attention to the linguistic and logical aspects of machine reading. The machine would require inference abilities beyond simple span extraction to answer the more advanced questions correctly.

The main contributions of this work are as follows. (1) We introduce the novel FLUD, the first large-scale, open-domain machine reading comprehension dataset with TOCFL Band C level questions and accompanying audio recordings. (2) To the best of our knowledge, it is the first corpus that considers deeper language understanding techniques common in human language and communication. (3) For national technology policy, it is also the most extensive Taiwanese Mandarin speech QA corpus that can enrich the development of accent-robust automatic speech recognition systems. (4) We conduct an extensive human evaluation of the quality and validity of the dataset's content and compare results from machine-learning models, providing us with a general understanding of current AI technologies used to solve these types of tasks.

2. Related Work

Many new large-scale reading comprehension datasets promote the rapid development of machine-learning models that can answer factual questions based on the provided information. This trend is pioneered by a notable breakthrough in the formulationof the Stanford Question Answering Dataset, or SQuAD (Rajpurkar et al., 2016). The innovation of this dataset lies in how the questions are composed; namely, each answer can be extracted in a reference article collected from Wikipedia. This dataset includes over 100,000 question-answer pairs along with their corresponding articles, which is much larger than previous work, e.g., MCTest (Richardson et al., (2013)). In addition, it is the first corpus that does not come in the form of multiple-choice questions, increasing the difficulty for a machine-learning model to guess the answer based on simple similarity measures between questions and choices. Another QA dataset published around the same time is the MS MARCO dataset (Nguyen et al. (2016)), which contains the same number of question-answer pairs, but with much more reference paragraphs (over 1 million). They are collected from user queries in the Bing search engine, and the answers have been composed by humans. TriviaQA (Joshi et al. 2017) studied this trend and made a few improvements. The 650,000 question-answer pairs and their corresponding document are collected from various trivia competition participants. Thus, they are created naturally, without reference documents. Then, the evidence articles are gathered from the Internet and Wikipedia. As a result, these questions are claimed to be more realistic and organic. The comprehension and inference of various sources, such as news, encyclopedias, and social media, is required to answer the questions in TriviaQA correctly. Subsequently, more QA datasets emerge, (e.g., HotpotQA Yang et al. (2018), ReCoRD Zhang et al. (2018), Cosmos QA Huang et al. (2019), and TYDI QA Clark et al. (2020)).

Almost all existing large-scale QA datasets consist of English content, whereas two notable Chinese corpora have been proposed, i.e., DuReader (He et al. (2018)) and DRCD (Shao et al. (2018)). The former includes simplified Chinese QA and the latter Traditional Chinese. DuReader contains questions from search engines as well as online forums, and the answers are generated by humans. There are 200,000 questions, 420,000 answers, and one million reference documents. DRCD stands for Delta Reading Comprehension Dataset, as constructed by Delta Electronics, Inc. It is an open-domain QA dataset, including over 10,000 paragraphs, 2,000 documents from Wikipedia, and over 30,000 questions. Similar to SQuAD and other extractive QA, the answer to each question can be extracted entirely from the accompanying paragraph and ismanually created by humans. The newly constructed FLUD stands out from previous research in several aspects. First, multiple logical inference techniques or linguistic knowledge are required to answer a question correctly. The answer to a question often cannot be directly extracted from the document. Second, open-ended essay questions can also be answered using the accompanying document apart from questions with exact answers. In addition, many multiple-choice

[1] The Test of Chinese as a Foreign Language (TOCFL) is a set of standardized language proficiency tests developed for non-native speakers of Chinese.

questions and task-oriented dialogue scripts are also included, further increasing the diversity of this dataset. Lastly, there are advanced questions where the answer covers multiple, noncontinuous spans of the reference article. These characteristics greatly enhance the level of understanding required to respond correctly to the questions.

3. Methodology

In this section, we describe how this corpus is constructed, including the collection of documents from various sources and composing the questions. Notably, The primary aims of building this dataset are to boost the development of machine reading comprehension and to be able to release it to the general public. Therefore, special attention is paid to assert the copyright of this data and the difficulty of the questions.

3.1. Data Collection

There are three parts in FLUD, namely, multiple-choice QA, short-answer QA, and dialogues. In this section, we introduce in detail the construction of each part.

3.1.1. Multiple-choice Questions

The first section of FLUD contains multiple-choice questions and their corresponding articles. There are 1,000 Band C and approximately 14,000 Band B level questions in this portion, respectively. According to TOCFL, the Band C level is designed for the following types of learners:

1.Non-native Chinese speakers at the advanced level.
2.Studied Chinese for over 960 hours in Taiwan or for 1,920 hours in non-Chinese-speaking regions.
3.Familiar with about 8,000 Chinese vocabulary terms.

Taking content diversity into consideration, we collect articles from the Science Development magazine and Chinese literary classics. The question domains include festivals, cuisines, traveling, music, sport, leisure, social and cultural topics in Taiwan, classical literature, Internet phenomenon, environment, and news in multichoice questions.

Nevertheless, collecting high-quality speech data that corresponds to the above content is time-consuming. In order to solve this problem, we inquire organizations such as radio stations to authorize us with the right to use their programs along with raw data from radio shows. For example, the Police Broadcasting Service (PBS) provideddaily news in 2018,and the National Education Radio contributed archived programs from 2012 to 2017. Moreover, we recruited more than 50 speakers to participate in the speech recording of the remaining questions. See Table 1for the sources, data type, and data for matfor this portion of the FLUD dataset.

After acquiring raw data, either in text or speech,we perform data cleaning for further use. We recruit annotators to carefully type and check every word while listening to the radio program. Then, they are asked to generate one to five questions from an article. Each question has four options and one answer. Answers to each question could be contained inthe article, while others may not. One example of a multiple-choice question set is listed in Table 2. Note that the exact wordsof the answer to this question are not present in the reference article.

Source	Data type	Format
National Education Radio (NER)	text and speech	Mono,16kHz, 16 bits PCM, *.wav
Police Broadcasting Service (PBS)	text and speech	Mono,16kHz, 16 bits PCM, *.wav
Science development (online magazine)	text and speech	Mono,16kHz, 16 bits PCM, *.wav
Chinese literary classics	text and speech	Mono,16kHz, 16 bits PCM, *.wav

Table 1. Source and data types of multiple-choice questions

3.1.2. Short-Answer Questions

We also utilize multiple sources for the second part of FLUD when collecting the short-answer and essay questions. The first major source is the Traditional Chinese version of Wikipedia (shortened as WTC hereafter). We gathered over 1,000 articles from WTC dumped on Aug. 20, 2019. Another source is from all elementary school textbooks published by the Taiwanese government before 1980 since they are open to the public. Next, we collected news articles from online news sites. The last source is the archive of public announcements made by the Taiwanese government. After collecting and cleaning the raw textual data, the question-answer pairs are manually created by three curators. In order to ensure the quality and accuracy of the answer(s), they are cross-checked by all curators. In the end, we compiled a training set of 746 question-answer pairs and over 100 related documents. Each question is assigned 1 or 2 points according to its complexity and difficulty.

More specifically, we design these questions so that they can cover a wider range of NLP techniques than straightforward span extraction. In order to answer theshort-answer questions correctly, the human/machine requires multiple types of inference strategies. The eight categories includel exical definition, enumeration, temporal relation, spatial relation, quantity, entailment, causal relation, and essay questions. For the reader's reference, the original Chinese termsare "字義, 列舉, 時間關係, 空間關係, 數量, 蘊含, 因果, 申論," respectively. There are also combinatory questions that require more than one of these types of methods. Detailed descriptions and some examples are as follows.

- Lexical definition: Using synonymy, hyponymy, antonymy, and definition information to understand the question. For instance, the question "What is the lastname of the U.S. President?" requires the definition of "last name."

Paragraph
臺灣有水果王國之美稱，一年四季都可以嘗到不同的水果，因為有多變化的地形，而且氣候舒適。春季有梅子、李子、琵琶，吃在嘴裡甜在心裡；夏季出產消暑解渴的西瓜、芒果、荔枝等，趕走悶熱的夏天；秋天正好吃柚子過中秋；冬季可以吃柑橘類：金棗、柑橘、柳丁酸甜滋味好像戀愛的感覺；一年四季更有鳳梨、蓮霧、木瓜等水果可以品嚐，臺灣的水果真好吃，如果來台灣旅遊，一定要試試看當季的水果。 Taiwan is known as the Kingdom of Fruits. Different fruits can be tasted all year round because of the changing terrain and the comfortable climate. In the spring, there are plums, apricots, and loquat, etc., which you could feel the sweet in your heart when tasting. In summer, there are thirst-quenching watermelons, mangoes, lychees, etc. to get rid of the sultry summer. In autumn, you can eat pomelo for mid-autumn festival. In winter, you can eat citrus: kumquat, citrus, orange, etc., the sweet and sour taste makes people feel that they are in love. There are pineapple, wax apple, papaya and other fruits that can be tasted throughout the year. Fruits in Taiwan are really delicious. If you come to Taiwan, you must try fruits in season
Question
請問文中的當季是什麼意思？ What does "in season" mean in the text?
Paragraph
1. 主要的季節 Main season 2. 符合生產的季節 Growing season 3. 豐富的水果種類 Rich variety
Paragraph
2. 符合生產的季節 Growing season

Table 2. An example of multiple-choice QA, along with the reference article in FLUD, with English translations

- Enumeration: The model must selectively extract multiple spans from the reference document. For example, we provide an article regarding COVID-19 and ask the question, "What are the symptoms of COVID-19?" The answers may be scattered over different parts of the reference article.
- Temporal relation: The model must know the meaning of, e.g., "next month," "last year," regarding the current or a specific day. For example, the reference article can be about the stock price of company X separated by year. We may ask: "What was the value of company X last year?"
- Spatial relation: The use of relations such as "on top of" and "within" to find the answer. For example, we can include a news article regarding an earthquake and the city's name in which the center is located. We may ask: "In what province did the earthquake originate?"
- Quantity: The ability to count or compare quantity. For example, there is an article about the number of deaths and hospitalized people due to COVID-19 and SARS-related diseases. We can then ask: "How many people were affected?" or "Which disease is more lethal?"
- Entailment: Understanding the logical relations betweenevents or entities, such as "A contradicts with B" or "A entails B." For instance, given the information "X is married to Y," one can answer, "Is Y a family member of X?"
- Causal relation: Utilizing the cause-effect relations in the document. For example, given the sentence "The disease can spread to birds, pigs, and humans." One can answer thequestion, "Can this disease affect multiple species?"
- Essay questions: the answer consists of a longer segment, often more than one sentence. It can be thought of as a form of summarization. There can be more than one correct response, as long as they correspond to the content ofthe reference article.

One example of a multiple-choice question set is shown in Table 3.

3.1.3. Task-oriented Dialogue

Lastly, FLUD also incorporates conversation scripts for 20 tasks. We compose 20 scripts, each targeting a specific task, as the training data of task-oriented conversational agents. They include train ticket reservations, purchasing home appliances, ordering drinks, travel recommendations, foreign exchange, savings account, restaurant reservations, sports game ticket reservations, and car rentals. We compose spreadsheets for each with important information that is required for their completion. Notably, there are situations where more than one client may contact the agent for some of these tasks, or the same client may call multiple times to change their previous request. We believe that these scenarios are more realistic and useful when training a task-oriented conversational machine.

Paragraph
國產毛豆去年外銷產值創下28年來的新高，外銷量3萬7520公噸，貿易額高達8118萬美元（約新台幣24.5億元），並促成7家冷凍食品公司於國內建新廠或擴廠，共投資計30億元，創造上千個就業機會，堪稱是「台灣綠金」！「毛豆」即未成熟且呈青綠色的食用大豆，全株的鮮莢80%達飽滿時，此時豆莢呈綠色帶有茸毛，故名為「毛豆」，又稱「菜用大豆」，日本稱為「枝豆」。一般豆類含有棉籽糖，容易引起脹氣。但毛豆的棉籽糖含量少，卻有豐富的鉀，可以改善因為缺乏鉀離子造成的倦怠和食欲下降，常做為開胃菜；它比一般豆類有更多優質蛋白質，被稱作植物肉，很適合素食者補充營養。農委會高雄區農業改良場長戴順發指出，近10年以來，臺灣的毛豆產業在周國隆先生和高、屏地區農民的通力合作下，除銷往日本占85.4%外，也銷往美國、加拿大等24個國家。戴順發說明，我國冷凍毛豆產品的產值，在日本市占率達到44.8%，分別是競爭對手泰國、中國的1.63、1.91倍，平均每公斤價格為250日圓，更較中國的189日圓31.9%。戴順發強調，高雄場領航品種研發，推出「高雄9號-綠晶」及「高雄11號香蜜茶豆」等高產毛豆品種授權產業界應用，同時為保護智慧財產權，除了申請國內品種權外，也向日本申請品種權，讓我國毛豆以創新研發勝出，跳脫市場削價競爭，也證明持續推動產業升級，是永續經營的方針。 The domestic export value of domestic edamame reached a new 28–year high, with an external sales volume of 37,520metric tons and a trade volume of US$81.18 million (about NT$ 2.45 billion). It also helped 7 frozen food companies tobuild new plants or expand factories in Taiwan. With an investment of NT$ 3 billion and thousands of job opportunities, itis called "Taiwan Green Gold"! "Edamame" is the immature and greenish–colored edible soybean. When the fresh pods ofthe whole plant are 80% full, the pods are green with fuzz, so it is called "edamame", also known as "vegetable soybean",and it is called "edamame bean" in Japan. Generally, beans contain cottonseed sugar, which may cause flatulence.

Question
1. 毛豆是否為台灣的經濟作物? Is edamame a Taiwanese cash crop? 2. 有「台灣綠金」之稱的是哪一種植物? Which plant is called "Taiwan Green Gold"? 3. 毛豆從哪一年開始外銷的? From what year did the edamame export start? 4. 毛豆有哪些別名? What other names are there for edamame? 5. 吃了豆類造成脹氣的原因是哪一種成份? What is the cause of flatulence from eating beans? 6. 文本中指出毛豆有哪些營養成份? In the text, which nutrients are found in edamame? 7. 多吃毛豆是否有益健康? Is eating edamame healthy? 8. 台灣毛豆外銷至哪些地區? To which countries are Taiwanese edamame exported? 9. 台灣毛豆的產區集中於何處? Where is the production area of Taiwanese edamame concentrated? 10. 欲申請品種權專利的毛豆品種是哪兩項? What are the two varieties of edamame that have applied for various patents? 11. 毛豆成為「台灣綠金」的優勢有哪些面向? What are the advantages of edamame as the "Taiwan Green Gold"?

Table 3. An example article in FLUD, with English translations, as the question reference

4. Evaluation

In this section, we first depict the design and the results of the human evaluation of FLUD. Then, we provide the machine performances achieved in the open competitions in which this dataset was used.

4.1. Human Evaluation

A predefined test set is split from FLUD for the purpose of the competition. Therefore, we assess human performance on the test set as the bar for machine-learning models. The human evaluation is divided into three parts according to the question format (e.g., multiple-choice, short-answer, and chatbot). First, similar to the TOCFL online test, 10 participants are asked to complete an exam with 50 multichoice questions in the form of spoken content. The duration of the test is 70 minutes. When recruiting the participants, the minimum education level is set as college or graduate school. The result of the human performance evaluation on the test set with multiple-choice questions is 89.6% in terms of accuracy.

Second, we randomly selected 30 from 167 participants who volunteered to join the short-answer exam in our study. Those who want to join the exam should meet the criteria as follows:

- foreigners who passed TOCFL Band C,
- or native speakers with a senior high school degree.

All of them are instructed to complete 50 short-answer questions individually, in the form of an online survey. The distribution of the question's level of difficulty matches that of the competition for machines. In other words, the number of basic and advanced questions match those in the competition. In the end, the average score is 50.21 out of 80. Therefore, the accuracy is equal to 62.7%.

For the last part of the test, namely, task-oriented conversational agent, we recruit five participants to perform multiround dialogue. Notably, the competition organizerselects five as the final testing category out of the 20 provided training topics in FLUD, including travel recommendations, restaurant reservations, insurance, car rentals, and housing loans. Therefore, each of the five participants in this experiment corresponds to one of those categories. Meanwhile, the participants in this experiment must have at least one year of customer service experience. The human customer service score is evaluated by a committee of five judges identical to the machine competition. In the end, the human score is 18 out of 20, denoting an accuracy of 90%.

4.2. Performance of Machine-Learning Models

The performance of machines is evaluated in an open competition, where participants provide computer programs that receive input from the organizers' machine and return correct responses. In order to evaluate the machine-learning models, they first must define the scoring criteria. The competition is scored in textual form. For multiple-choice questions, the criterion is clear, namely, selecting the correct answer. The short-answer questions and multiturn dialogue tasks are reviewed and scored by the committee. The essence of the answering criteria is listed in Table 4, and the complete document can be found on the competition website.

Two open competitions using different parts of FLUD were held. During the first multiple-choice competition, a machine must answer 1,000 multichoice questions in 90 minutes. Interestingly, the organizer of this competition posed an extra challenge. The participating machines must first perform speech recognition since the input is in the form of wave (WAV) files. There is no limitation regarding the resources that they can utilize. In other words, the participants can make use of third-party speech recognizers and knowledge bases. In the end, the best performance in this competition is 53.7% in terms of accuracy.

For the other two tasks, the participants (machine-learning models) must answer 50 questions (short-answer and essay) and five task-oriented dialogues in this competition without external resources. The total score sums up to 100, in which the 50 questions account for 80 points. Recall that some advanced questions are worth 2 points and others 1 point. Moreover, essay questions and dialogue transcripts are evaluated by the committee—the finalist scores of this competition range from32.7 to 15.3 for the short-answer and essay questions. For dialogue agents, the scores are between 14 and 4. Finally, the total score among 10 finalists ranges from 36.7 to 4 out of 100. Below is the performance comparison between humans and machines.

Judgment	Criteria
Incorrect	Irrelevant answers, incomplete sentences,wrong sentence structure, unclear, and wrong meaning
Correct	Appropriate description, fluent sentence, high-quality content information, precise words, or extensive knowledge

Table 4. Scoring criteria for short-answer and essay questions in FLUD

	Multiple-choice	Short-answer	Task-oriented chatbot
Human Performance	89.6%	62.7%	90%
Best Performances of Machine-Learning Models	53.7%	40.8%	70%

Table 5. Comparison between the performances of humansand machines in three different tasks.

5. Conclusions

We introduce the FLUD, a new Traditional Chinese machine understanding dataset. FLUD is the largest dataset focusing on deeper linguistic knowledge and substantial amounts of spoken content. To the best of our knowledge, the dataset is also the first speech corpus published by a noncommercial entity that covers speakers of all ages using Taiwanese Mandarin. Notably, there are four unique features of FLUD. (1) It consists of four types of questions that can serve as a stepping stone to the development of machine comprehension models. (2)The background knowledge within this dataset contains vocabulary, common sense, and geographical information specific to the Taiwanese people/region. (3)The cultural diversity of this country is also reflected in the dataset. Specifically, the topics include festivals, cuisine, traveling, music, sports and leisure, social and cultural activity, classical literature, Internet phenomenon, environment, and news. (4) It is the largest Taiwanese Mandarin speech QA corpus that contains around 400 hours of recordings. (5) The level of language understanding required to answer the questions is very sophisticated, as evidenced by the comparison between the human evaluation and machine competition results. We envision that FLUD can significantly enhance the development of NLP and AI technologies worldwide.

Acknowledgement

First, I would like to acknowledge the support of the Science & Technology Policy Research and Information Center at Narlabs, particularly the former director-general Dr. Yuh-Jzer Joung, for giving us the opportunity to start the research. Many thanks are also owed to Dr. Teyi Chan, Dr. Jui-Shin Chang, and Professor Hung-yi Lee for their assistance and patience. Finally, we thank the National Center for High-performance Computing for providing computational and storage resources.

References

Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., & Palomaki, J. (2020) Tydi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics 8, 454-470.

He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T.,& Wang, H. (2018) Dureader: a Chinese machine reading comprehension datasetfromreal-world applications. Proceedings of the Workshop on Machine Reading for Question Answering,37-46.

Huang, L., Le Bras, R., Bhagavatula, C., & Choi, Y. (2019) Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. Proceedingsof the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP- IJCNLP), 2391-2401.

Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017) Trivia QA:A large scale distantlysupervisedchallengedatasetforreadingcomprehension.Proceedingsof the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),1601-1611.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., & Deng,L. (2016) MS Marco: A human generated machine reading comprehension dataset. In CoCo@NIPS.

Shao, C. C., Liu, T., Lai, Y., Tseng, Y., & Tsai, S. (2018) DRCD: a Chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen,W., Salakhutdinov, R., & Manning, C. D. (2018) Hotpot QA: A dataset for diverse, explainable multi-hop question answering. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2369-2380.

Zhang, S., Liu, X., Liu, J., Gao, J., Duh, K., & Van Durme, B. (2018) Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885.