A Method for Building a Commonsense Inference Dataset based on Basic Events

Kazumasa Omura **Daisuke Kawahara*** **Sadao Kurohashi**
Graduate School of Informatics, Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501, Japan
{omura, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

We present a scalable, low-bias, and low-cost method for building a commonsense inference dataset that combines automatic extraction from a corpus and crowdsourcing. Each problem is a multiple-choice question that asks contingency between basic events. We applied the proposed method to a Japanese corpus and acquired 104k problems. While humans can solve the resulting problems with high accuracy (88.9%), the accuracy of a high-performance transfer learning model is reasonably low (76.0%). We also confirmed through dataset analysis that the resulting dataset contains low bias. We released the datatset to facilitate language understanding research.[1]

1 Introduction

Along with the progress of deep learning, there have been many studies that consider task settings and build their datasets for training/evaluating language understanding ability by computers (Wang et al., 2019b,a).

Language understanding by computers requires two types of knowledge: knowledge of language (meaning of words, syntax, and so forth) and knowledge of our world and society beyond language.

The former problem of acquiring linguistic knowledge has been solved to a large extent by general-purpose language models, such as BERT (Devlin et al., 2019), which are pre-trained using a large corpus. It is now possible to represent the meaning of a word as a vector according to its context. Fine-tuning based on these vectors has made natural language inference, paraphrase recognition, and question answering without requiring deep inference as accurate as humans.

On the other hand, there are still many problems with acquiring knowledge beyond language. Actually it is open-ended, and we had better start with fundamental knowledge, i.e., commonsense. Still, it is not easy to focus on commonsense, guaranteeing some generality as commonsense.

There have been some approaches to guarantee such generality. SWAG (Zellers et al., 2018), for example, focuses on knowledge about daily events that can be visually perceived. This method greatly limits the range of commonsense that can be acquired. CommonsenseQA (Talmor et al., 2019) is based on the basic vocabulary that is covered by ConceptNet (Speer et al., 2017), which is one of the largest commonsense knowledge bases. This prevents the scalability, generating only 12k problems from the whole data of ConceptNet.

Another important point is that biases in building datasets must be reduced as much as possible. In the above two approaches, question or distractor sentences were created automatically or by crowdsourcing. This causes generation biases of language models or produces certain patterns (*annotation artifacts*) by crowdsourced writing (Gururangan et al., 2018).

We use a text corpus to solve these problems. We propose a method to build a commonsense inference dataset by extracting contingent pairs of basic event expressions (hereafter, **contingent basic event pairs**) from a corpus and verifying them by crowdsourcing. Basic event expressions (hereafter, **basic events**) are defined as expressions composed of high-frequency predicate-argument structures that are extracted from a corpus and aggregated by clustering according to their usages. Contingent basic event pairs are extracted by identifying contingency relations between basic events using discourse parsing.

For instance, the following contingent basic event pairs are acquired.

*Current affiliation is Waseda University.
[1]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JCID

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 2450–2460,
November 16–20, 2020.

I am hungry, so
a. I drink coffee.
b. I have a meal.
c. I sweat.
d. I get sleepy.

Table 1: A sample problem. The correct choice is bolded.

(1) a. I am hungry, so I have a meal.
b. If I have a meal, I get sleepy.
c. Since I am sleepy, I drink coffee.
d. If I exercise hard, I sweat.

Based on these contingent pairs, we can generate a **commonsense inference problem** by adopting latter events of other pairs as distractors, as shown in Table 1. Since the problem is based on basic events, it guarantees some generality as commonsense.

Since our method is based on automatic extraction from a corpus, it is scalable and the domain is not limited. In addition, there is no bias caused by crowdsourcing because we ask crowdworkers to just verify a sentence. Although the key idea of our proposed method is language-independent, we build a Japanese commonsense inference dataset in this study by exploiting existing resources.

The contributions of this paper are summarized as follows.

- We propose a scalable, low-bias, and low-cost method for building a commonsense inference dataset that combines automatic extraction from a corpus and crowdsourcing.
- We built a Japanese commonsense inference dataset from a web corpus of 715m sentences that consists of 104k multiple-choice questions.
- While humans can solve the resulting problems with high accuracy (88.9%), the accuracy of a high-performance transfer learning model is low (76.0%), which shows that there is a reasonable gap in commonsense inference ability. We also confirm that the resulting dataset contains low bias.

2 Related Work

Language resources for commonsense inference that have been built so far can be classified into knowledge bases and QA datasets.

Commonsense knowledge bases have been constructed by experts, crowdsourcing, and games with a purpose. Cyc (Lenat, 1995) and Open Mind Common Sense projects (Speer et al., 2017) collected various relations between entities and events. They include causal relations between events, but the number of these relations is not high. ATOMIC (Sap et al., 2019) is a knowledge base that is comprised of 877k if-then pairs of basic events. They collected these pairs using crowdsourcing based on frequent basic events extracted from several corpora. These fully manual or crowdsourcing approaches are costly and have a problem of scalability. Also, methods for incorporating such knowledge bases into an NLP model have been studied but have not been established yet.

Many QA datasets for commonsense inference have been built. They include COPA (Choice of Plausible Alternatives) (Roemmele et al., 2011), SWAG (Zellers et al., 2018), HellaSWAG (Zellers et al., 2019), and CommonsenseQA (Talmor et al., 2019). These datasets can be solved to some extent by machine comprehension models (Devlin et al., 2019) that have been rapidly improved. There have been also some approaches that transfer knowledge in such a dataset to downstream tasks using multi-task learning (Liu et al., 2019). We briefly introduce these datasets below.

COPA consists of 1,000 two-choice questions that ask a causal relation between two sentences. Each question provides a premise sentence and requires to choose its cause or ending sentence from two alternatives. This dataset was manually created for the purpose of evaluation and is too small to learn commonsense by computers.

SWAG is a commonsense inference dataset consisting of 113k multiple-choice questions that ask the most appropriate verb phrase following a given context. To guarantee generality as commonsense, questions were created from video captions, and thus the domain of the dataset is limited to the physical world. For each question, two consecutive sentences were extracted from a video caption, the first sentence and the subject of the second sentence compose a context, and the rest was regarded as a correct choice. Distractors were generated from a language model. To obtain high-quality distractors, SWAG removed those that are easily discriminated by an answer model. SWAG was solved by BERT with a similar accuracy to humans. This was attributed to biases that were embedded in

distractor sentences by an LSTM-based language model and detected by BERT (Zellers et al., 2019). They newly built HellaSwag using a better language model to make biases undetectable by BERT. However, the accuracy for solving HellaSwag is also approaching to human performance (Liu et al., 2020). The bias problem has not been solved yet.

CommonsenseQA is a commonsense inference dataset consisting of 12k multiple-choice questions based on the commonsense knowledge base, ConceptNet. A question is created by crowdsourcing based on a subgraph consisting of a source concept and three target concepts connected with the same relationship. A crowdworker creates a question sentence which includes the source concept and whose answer is only one of the target concepts. This method uses an existing resource and lacks scalability. In addition, because the load of creating question sentences is large for crowdworkers, they tend to use the same words and styles repeatedly, leading bias in question sentences.

3 A Method for Generating Commonsense Inference Problems

A commonsense inference problem consists of a context (question) and four choices. The question asks to choose the most appropriate choice following the context, as shown in Table 1.

These problems should be based on basic events to guarantee generality as commonsense. In addition, to guarantee scalability and reduce biases, we combine automatic extraction from a corpus and verification by crowdsourcing. Our method to generate commonsense inference problems consists of the following procedure (Figure 1).

1. Acquire basic events from high-frequency predicate-argument structures.

2. Apply discourse parsing to a corpus and extract event pairs that are recognized as having a contingency relation and composed of basic events.

3. Verify whether the extracted event pairs have a contingency relation by crowdsourcing and obtain contingent basic event pairs.

4. Generate commonsense inference problems by taking a correct choice from a contingent pair and selecting distractors from other event pairs.

Case frame	CS	Case fillers
kowasu (1) (injure)	*ga* 1756 *wo* 70135 *de* 3941	I 83, person 65, ... stomach 25643, body 17242, ... stress 297, eating 174, ...
kowasu (2) (destroy)	*ga* 502 *no* 10147 *wo* 18274	person 42, Japan 42, ... place 873, room 851, ... atmosphere 8140, image 3774, ...
...		

Table 2: Examples of Japanese case frames. CS denotes case slots, where *ga*, *wo*, *de*, and *no* mean nominative, accusative, instrumental, and genitive, respectively. The number following case or a case filler represents its frequency. Examples are expressed only in English for space limitation.

We describe the details of each step in the following subsections.

3.1 Acquisition of Basic Events

Basic events are defined as expressions composed of high-frequency predicate-argument structures that are extracted from a corpus and aggregated by clustering according to their usages. As a source of basic events, we employ *case frames* (Kawahara et al., 2014) that are automatically constructed by clustering predicate-argument structures.

In the case frame data, each predicate has multiple case frames distinguished according to their usages. Each case frame consists of multiple case slots, and each case slot contains possible case fillers. Table 2 shows some examples of Japanese case frames.

In this study, we extract high-frequency predicate-argument structures from case frames as basic events. First, from the case frame data, the top α predicates in active voice are obtained. For each predicate, frequent case frames, case slots, and case fillers are selected until the cumulative sum of frequencies reaches $\beta\%$, $\gamma\%$, and $\delta\%$, respectively. For example, case frames are selected until covering $\beta\%$ of the frequency of a target predicate. These thresholds are empirically set according to a target language.

Table 3 shows some examples of basic events acquired from Japanese case frames. The parameters for Japanese basic events are described in Section 4.

3.2 Automatic Extraction of Contingent Basic Event Pairs

We apply dependency and discourse parsing to a text corpus and extract event pairs connected with both dependency and contingency relations.

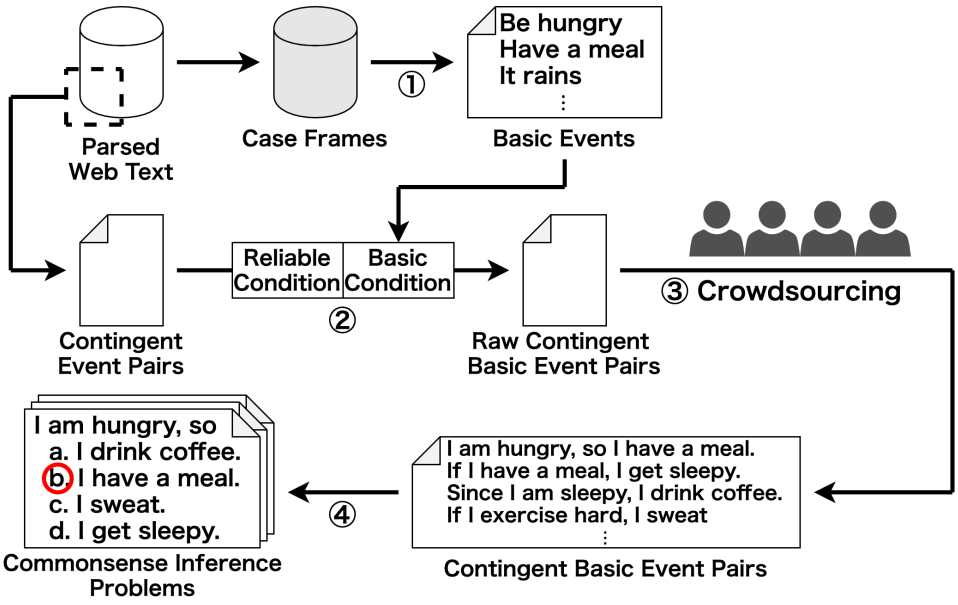

Figure 1: The overview of our proposed method.

Case frame	CS	Case fillers
kowasu (1) (injure)	*wo*	stomach, body
kowasu (2) (destroy)	*no*	place, room, ...
	wo	atmosphere, image
torikaeru (replace)	*wo*	door, glass, ...
hatudou (activate)	*ga*	effect

Table 3: Examples of Japanese basic events.

The contingency relation between events should be expressed by an explicit discourse marker and be a causal or conditional relation, corresponding to "CONTINGENCY:Cause" or "CONTINGENCY:Condition" in the Penn Discourse Treebank (Prasad et al., 2008).

To select highly reliable parts from analysis results and to extract only general event pairs as commonsense, we keep event pairs satisfying the following conditions. Here, we call the first event that represents a cause or reason **former event** and the second event **latter event**.

Reliable The former and latter events are unambiguously connected.

In the case that only two clauses exist in a sentence, there is no ambiguity. In the case that more than two clauses exist in a sentence, we extract a reliable part based on language-dependent criteria.

Basic Both the former and latter events are composed of a basic event.

This condition can be applied in a straightforward way, but we need to take care of the case that an argument in the latter event is pronominalized or omitted. If the latter event does not have an explicit argument, we recover it with any of the arguments in the former event and examine whether the recovered latter event is composed of a basic event.

For example, consider the event pair "the glass breaks on impact → I replace it". In this case, we generate recovered latter events "I replace the glass" and "I replace impact" by substituting an argument in the former event for "it". Then, we examine whether either of them is composed of a basic event and extract this event pair because "replace glass" is a basic event as shown in Table 3.

Finally, the following post-processes are performed so that crowdworkers in the next step can accurately judge event pairs.

- To exclude event pairs that are less eventful or contain web-specific functional expressions, the frequency of basic events included in the obtained event pairs is counted, and event pairs that contain one of high-frequency basic events are excluded. For example, "問題がない (have no problem)" and "情報が満載 (have much information)" are detected as high-frequency meaningless events in Japanese.

- Event pairs that contain demonstratives or unknown words are excluded.

3.3 Verification of Contingent Basic Event Pairs through Crowdsourcing

We select contingent basic event pairs from the extracted basic event pairs using crowdsourcing. We ask crowdworkers to select one of the following two alternatives for each event pair.

1. A is a cause or reason of B.
2. Other relation or no relation.

Here, "A" denotes the former event, and "B" denotes the latter event.

We ask multiple workers to evaluate each event pair and adopt the evaluation that half or more of the workers agree. We finally obtain event pairs whose aggregated evaluation is "A is a cause or reason of B" as contingent basic event pairs.

3.4 Generation of Commonsense Inference Problems

We generate commonsense inference problems from the obtained contingent basic event pairs. We regard the former event as a context (question) and the latter event as a correct choice. Distractors are automatically selected from the latter events of other event pairs.

In general, highly similar distractors to the correct choice are not distinguishable even by humans. Meanwhile, dissimilar distractors can be easily distinguished by machines. We select moderately similar distractors under the following conditions.

Choice-Similarity The similarity between the correct choice and a candidate latter event is in a range, $\text{RANGE}_{\text{choice}}$.

This similarity is calculated using the cosine similarity between vectors of (latter) events. This vector is defined as an average vector of content words contained in an event.

Context-Similarity The similarity between the context and the former event of a candidate latter event is in a range, $\text{RANGE}_{\text{context}}$.

This similarity is calculated in the same way as the condition **Choice-Similarity**.

To improve the appearance of problems, we select latter events whose ratio of the number of words against the correct choice is in a range, $\text{RANGE}_{\text{length}}$.[2]

If more than three distractors are obtained, we randomly select three out of them. If less than three distractors are obtained, we do not generate a problem from the contingent basic event pair.

4 Building a Japanese Commonsense Inference Dataset

We built a Japanese commonsense inference dataset using the method described in Section 3.

Acquisition of basic events
We extracted Japanese basic events from the Kyoto University case frames[3], which had been constructed from 10 billion web sentences. We set the thresholds α, β, γ, and δ to 5,000, 75, 50, and 50, respectively. As a result, we obtained 28,642 basic events. Examples of the obtained basic events are shown in Table 3.

Automatic extraction of contingent basic event pairs
We automatically extracted contingent basic event pairs from a Japanese web corpus consisting of approximately 715 million sentences. We used the Japanese parser, KNP[4] to extract event pairs from the corpus. KNP does dependency parsing and also labels explicit discourse relations between clauses (events). As a result, approximately 85 million contingent basic event pairs were extracted.

Next, to extract highly reliable basic event pairs, the **Reliable** and **Basic** conditions were applied. For the **Reliable** condition, if there are more than two clauses in a sentence, we extract only the last two clauses because in Japanese the dependency goes from left to right.

Finally, we performed the post-processes to extract 164,910 contingent basic event pairs. The detailed statistics are listed in Table 4.

To investigate the effectiveness of the **Basic** condition, we randomly selected 100 event pairs from "+Reliable" and "+Reliable+Basic" in Table 4, and manually evaluated them. For convenience, we name each set of the selected event pairs "R" and

[2] As a result of our preliminary experiment, we found that this condition did not affect the model performance. Hence, we do not investigate this condition.

[3] https://www.gsk.or.jp/catalog/gsk2018-b

[4] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?KNP

Item	Number
sentences	714,605,164
contingent event pairs	85,357,299
+Reliable	51,904,745
+Reliable+Basic	517,321
+post-processing	164,910

Table 4: Detailed statistics about extraction of event pairs. For example, the number of "+Reliable" shows the number of contingent event pairs that satisfy the **Reliable** condition.

"RB", respectively. As a result of the manual evaluation, 47 event pairs in "R" and 76 event pairs in "RB" were judged as understandable with commonsense. Here is an example in "R" that would be excluded by the **Basic** condition: "サイクロンを発動すると→破壊できる (activate the cyclone→we can destroy)". "Activate the cyclone" is not a basic event as shown in Table 3, which is a domain-specific expression especially used in fiction. By the **Basic** condition, we can remove such non-general event pairs. Thus, we can see that the **Basic** condition is effective in acquiring general knowledge at the level of commonsense.

Verification of contingent basic event pairs through crowdsourcing

Next, we selected contingent basic event pairs from the extracted event pairs using crowdsourcing. We used the crowdsourcing service, Yahoo! Crowdsourcing[5]. A crowdworker was presented with 17 questions (event pairs) per task, each of which asked to choose one from the two alternatives. Two of the 17 event pairs were check questions with a hidden ground truth, and the answers of crowdworkers who mistakenly judged these event pairs were excluded. Each event pair was verified by four crowdworkers, and we selected the event pairs two or more of whose evaluations are "A is a cause or reason of B".

As a result of crowdsourcing, 104,266 contingent basic event pairs were selected from 164,910 pairs, which means that approximately one-third of pairs were removed. This ratio roughly corresponds to the result of the above investigation on the effectiveness of the **Basic** condition. The total cost of crowdsourcing was 484,000 JPY (4,495 USD), and the cost per problem was 4.7 JPY (4.5 cents).

[5] https://crowdsourcing.yahoo.co.jp/

Train	Development	Test
83,127	10,228	10,291

Table 5: Statistics of the dataset.

Generation of commonsense inference problems

Finally, we generated commonsense inference problems from the acquired contingent basic event pairs. The similarity range, $\text{RANGE}_{\text{choice}}$, in the condition **Choice-Similarity** was set to the range of 0.4 to 0.6, and the similarity range, $\text{RANGE}_{\text{context}}$, in the condition **Context-Similarity** was set to the range of 0.5 to 0.7. We set $\text{RANGE}_{\text{context}}$ slightly higher than $\text{RANGE}_{\text{choice}}$ because **Context-Similarity** controls the similarity to the correct choice more indirectly than **Choice-Similarity**. To calculate the similarity between events, we used word vectors that were induced from 200 million sentences of the Japanese web corpus using word2vec[6]. The length range, $\text{RANGE}_{\text{length}}$, was set to the range of 0.5 to 2.0.

As a result, 103,907 problems were generated from the 104,266 contingent basic event pairs. Table 7 shows examples of the obtained problems with BERT's predictions (described in Section 5.1). On this default setting, the mean and median numbers of the eligible candidates before finally selecting three were 3,459 and 1,355, respectively.

To create a standard split of the obtained problems, we split the problems into train/development/test sets with the ratio 8:1:1. We performed the split in the way that both the train set and the development/test set do not contain the problems generated from the identical "seed". The term "seed" refers to a pair of basic events that compose former and latter events in a contingent basic event pair. For example, the seed of the top left example in Table 7 is a pair of "装置が故障 (a device breaks)" and "装置を交換 (replace a device)". In addition, we removed some problems in the development/test sets so that there are no duplicate pairs of a context and a distractor between the train set and the development/test sets. The statistics of the resulting dataset are listed in Table 5.

Investigation of human accuracy

To investigate the accuracy of human answers, we randomly sampled 1,500 problems and collected answers from five crowdworkers for each problem.

[6] https://code.google.com/archive/p/word2vec/

We prepared three sets of 500 problems and did crowdsourcing on different dates to be solved by different sets of crowdworkers. As a result, the average accuracy of individual crowdworkers was 83.8% and that of the answers aggregated by majority voting was 88.9%.

5 Experiments

We conducted experiments to investigate the performance of a transfer learning model on the constructed commonsense inference dataset.

5.1 Model

We used BERT (Devlin et al., 2019) as a transfer learning model for our experiments. BERT achieved high performance on various benchmark tasks including natural language inference and question answering. For pre-training, the model solves a masked language modeling task and a next sentence prediction task simultaneously to obtain context-aware word representations. To apply this model to each downstream task, a layer is added on top of the output, and all the parameters are fine-tuned on the task.

In our experiments, we input pairs of a context and a choice separated by special tokens following the previous work (Talmor et al., 2019). For example, the context "お腹 が 空いた ので (I am hungry, so)" and the choice "ご飯 を 食べる (I have a meal)" would become "[CLS] お腹 ... [SEP] ご飯 ... [SEP]". The hidden representation of each [CLS] token is converted to a score through a linear layer, and the choice with the highest value is selected as an answer.

We define the objective function as follows.

$$L = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(\mathbf{w}^T\mathbf{c}_{kj})}{\sum_{i=1}^{4}\exp(\mathbf{w}^T\mathbf{c}_{ki})} \quad (1)$$

where N is a batch size, $\mathbf{w}$ is the parameters in a linear layer, j is the index of a correct choice among $1 \cdots 4$, and $\mathbf{c}_{ki}$ is the hidden representation of each [CLS] token.

We adopted $\text{BERT}_{\text{LARGE}}$ as a BERT model. We used the Japanese pre-trained $\text{BERT}_{\text{LARGE}}$ WWM model[7], which performed pre-training using 18 million sentences of Japanese Wikipedia with whole word masking. We fine-tuned the pre-trained model for 3 epochs. We used the following hyper-parameters: a batch size of 8[8], a learning rate of 2e-5, and maximum sequence length of 128.

Model		Accuracy
Chance		0.250
$\text{BERT}_{\text{LARGE}}$		0.760
Human	1 worker	0.838
	5 workers	0.889

Table 6: Performance of $\text{BERT}_{\text{LARGE}}$ and humans.

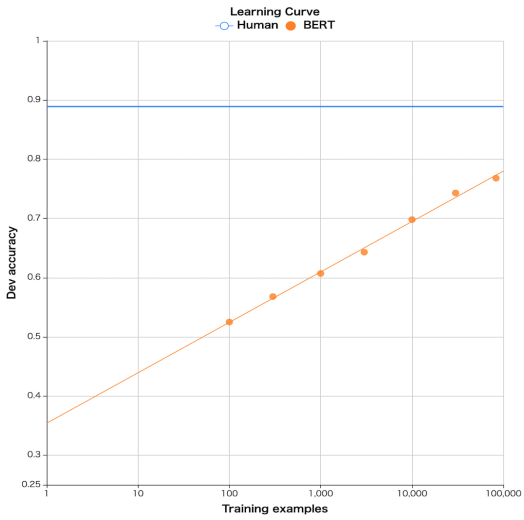

Figure 2: Learning curve of the BERT model on the development set.

5.2 Experimental Results

We evaluated the model performance with accuracy. The $\text{BERT}_{\text{LARGE}}$ model achieved an accuracy of 0.760, as shown in Table 6. We can see that there is a reasonable gap between the BERT model and the human performance.

Figure 2 shows the learning curve of the BERT model on the development set. We can expect by extrapolation that approximately 1.9 million training examples are required to achieve human performance, which is not practical. It is meaningful to develop better models to solve this dataset toward the human accuracy.

5.3 Analysis

We briefly analyze the results of the BERT model. Table 7 shows some examples that the BERT model answered correctly and incorrectly. As can be seen from the top left example, lexical overlap between a context and a choice is a clue to solve the problem. There are some noticeable examples that the BERT model answered incorrectly as a result of overemphasizing this.

[7] http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT%E6%97%A5%E6%9C%AC%E8%AA%9EPretrained%E3%83%A2%E3%83%87%E3%83%AB

[8] Each batch corresponds to one problem, that is, consists of four input sequences.

correct	テープ装置が故障したので (Since a tape device broke,) a. 黒い方がシャープに見える$_{-14.5}$ (the black one looks sharper) b. 購入を決める$_{-4.4}$ (I decide to buy) **c. テープ装置を交換します$_{13.7}$** **(I replace the tape device)** d. 撮影の幅が大きく広がる$_{-14.5}$ (you can add variety to your photography)	今はなにしろ９時に寝ないといけないので、 (Since I have to go to bed at nine anyway,) a. 進行をできるだけ抑えるための治療が必要だ$_{-14.4}$ (I need treatment to prevent disease progression as much as possible) b. わかりやすく教えていただけましたら助かります$_{-14.0}$ (I'd be grateful if you would kindly explain it) **c. 敢えて面白そうな番組も見ないようにしています$_{2.4}$** **(I try not to watch an interesting TV show)** d. 供給も可能かもしれません$_{-14.6}$ (I may be able to provide it)
incorrect	仕事辞めたら (If you quit a job, then) **a. 生活は大変だ$_{-6.4}$** **(you lead a hard life)** b. 仕事でいっぱいいっぱいだ$_{-6.1}$ (you are exhausted from work) c. 勉強がはかどるはずだ$_{-9.3}$ (you must make progress in your studies) d. ますます犯罪が増えるだろう$_{-10.1}$ (the number of crimes will increase)	ウナギよりも脂が少ないので (Since it is less fatty than eel,) **a. あっさりとした味が楽しめます$_{4.7}$** **(you can enjoy a light taste)** b. 今回は、お塩は使用しませんでした$_{10.6}$ (I did not use salt this time) c. フライドポテトみたいな感じで美味しい$_{9.1}$ (it tastes good like french fries) d. ミネラルや水分の摂取など、食事面の配慮も必要だ$_{-9.4}$ (you need to take care of your nutrition, e.g. minerals and moisture)

Table 7: Examples that the BERT model answered correctly and incorrectly. The correct choice is bolded. If the BERT model answered incorrectly, its prediction is highlighted in red. The number at the end of each choice represents an output score.

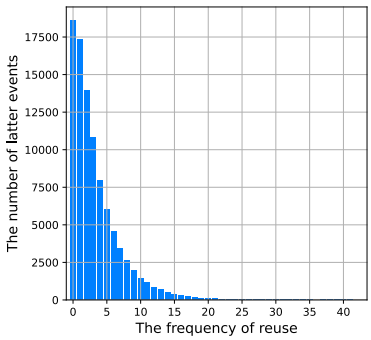

Figure 3: Counts of how many times each latter event is used as a distractor.

6 Investigation of the Dataset

6.1 Investigation of Biases

Several studies have reported that, due to unintended biases in a dataset, many problems can be solved by just observing a part of question sentences (Gururangan et al., 2018; Zellers et al., 2019). To investigate the existence of bias in our dataset, we measured model performance when we input only the choices by omitting their context during fine-tuning and inference phases. In this investigation, we used the same model and hyperparameters as described in Section 5.1.

As a result, the BERT$_{\text{LARGE}}$ model achieved an accuracy of 41.2%. Compared with the result in Section 5.2, the performance is significantly low, which indicates that the constructed dataset contains low bias. To investigate the result that the performance without the context (41.2%) is a bit higher than the chance rate (25%), we counted how many times each latter event is used as a distractor. Figure 3 shows the result of counting. We speculate that some latter events are frequently reused and thus can be easily judged as incorrect by the BERT model. We will tackle this problem in the future to further lower the bias.

6.2 Investigation of the Conditions on Selecting Distractors

We investigated how the conditions on selecting distractors affect the quality of the dataset. Specifically, we built datasets by removing the upper or lower bounds of each similarity range, RANGE$_{\text{choice}}$ and RANGE$_{\text{context}}$, and evaluated model and human performances on each dataset. We evaluated model performance on each development set using the same model settings as described in Section 5.1. We calculated human performance in the same way as described in Section 4.

Table 8 shows the result of this investigation. This result indicated the effectiveness of the upper and lower bounds. Specifically, by removing the upper bound, some problems contained distractors that were highly similar to the correct choice, and thus both the model and humans could not solve

RANGE		$\text{BERT}_{\text{LARGE}}$	Human
choice	context		
(0.4, 0.6)	(0.5, 0.7)	0.768	0.889 (0.838)
(0.4, 1.0)	(0.5, 0.7)	0.727	0.822 (0.788)
(0.4, 0.6)	(0.5, 1.0)	0.730	0.818 (0.777)
(-1.0, 0.6)	(0.5, 0.7)	0.767	0.887 (0.839)
(0.4, 0.6)	(-1.0, 0.7)	0.846	0.928 (0.888)

Table 8: Results of investigation of the conditions on selecting distractors. The numbers in parentheses at the rightmost column represent average accuracies of individual crowdworkers.

them. By removing the lower bound, the relatedness between a context and distractors decreased, and thus the generated problems became easy to solve especially for the model. Accordingly, it is important to select moderately similar distractors.

7 Conclusion

In this paper, we proposed a scalable, low-bias, and low-cost method for building a commonsense inference dataset that combines automatic extraction from a corpus and crowdsourcing. Each problem is a multiple-choice question that asks contingency between basic events. We applied the proposed method to a Japanese web corpus and acquired 103,907 problems. While the human accuracy was high (88.9%), the $\text{BERT}_{\text{LARGE}}$ accuracy was reasonably low (76.0%). We also confirmed that the dataset contained low bias, and thus it can be used as a good benchmark for language understanding research.

In the future, we will make a model learn commonsense with the obtained dataset and consider applying it to semantic tasks, such as anaphora resolution and discourse parsing.

For commonsense acquisition from text, there is a problem that every commonsense is not written in text because of reporting bias (Gordon and Van Durme, 2013). To acquire a wider range of commonsense, it is possible to combine our method with other methods based on physical world resources, such as video captions used in SWAG.

Acknowledgments

We thank anonymous reviewers for their valuable comments. This work was supported by the Japan Kanji Aptitude Testing Foundation.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. 2014. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 58–67, Gothenburg, Sweden. Association for Computational Linguistics.

Douglas B. Lenat. 1995. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Commun. ACM*, 38(11):33–38.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. 2020. Adversarial Training for Large Neural Language Models. *arXiv*, abs/2004.08994.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Logical Formalizations of Commonsense Reasoning,*

Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011. AAAI.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 3266–3280. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Screenshots of the crowdsourcing tasks

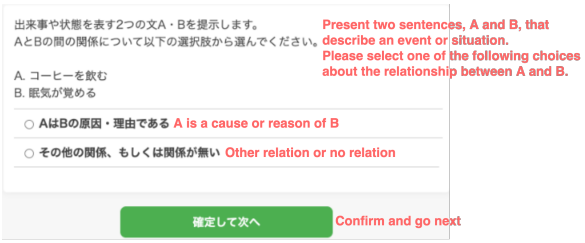

Figure 4: The screenshot of verification of contingent basic event pairs.

Figure 5: The screenshot of investigation of human accuracy.