Improving Commonsense Contingent Reasoning by Pseudo-data and its Application to the Related Tasks

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

Abstract

Contingent reasoning is one of the essential abilities in natural language understanding, and many language resources annotated with contingent relations have been constructed. However, despite the recent advances in deep learning, the task of contingent reasoning is still difficult for computers. In this study, we focus on the reasoning of contingent relation between basic events. Based on the existing data construction method, we automatically generate large-scale pseudo-problems and incorporate the generated data into training. We also investigate the generality of contingent knowledge through quantitative evaluation by performing transfer learning on the related tasks: discourse relation analysis, the Japanese Winograd Schema Challenge, and the JCommonsenseQA. The experimental results show the effectiveness of utilizing pseudo-problems for both the commonsense contingent reasoning task and the related tasks, which suggests the importance of contingent reasoning.

1 Introduction

Contingency is the relation between two events, one being an action or state and the other being likely to happen after it. We humans reason contingent relation between events on a daily basis. For instance, when we read text, we unconsciously infer what happens next to deepen our understanding. In conversations, we guess the next topic from the utterance of the opponent to make a contextual and natural response. Thus, the ability to reason contingent relation between events is essential when it comes to natural language understanding (NLU).

Recently, many studies have built language resources for contingent reasoning (Roemmele et al., 2011; Mostafazadeh et al., 2016; Zellers et al., 2018; Sap et al., 2019a; Hwang et al., 2021). These resources focus on basic events and evaluate some kind of commonsense reasoning ability. Although the fundamental linguistic capabilities of computers, such as question answering, have greatly improved with progress in deep learning, several studies have empirically demonstrated they still have difficulty in commonsense reasoning (Talmor et al., 2019; Sap et al., 2019b; Talmor et al., 2021).

I'm hungry, so
a. I'm gonna be absent from school.
b. I refrain from strenuous exercise.
c. I have a meal at a family restaurant.
d. I leave home.

Figure 1: Example from KUCI (English translated version). KUCI is a Japanese QA dataset containing 104k multiple-choice questions regarding contingent relation between basic events. The correct choice is bolded.

In this study, we aim at two objectives: to improve commonsense contingent reasoning and to investigate the effects of learning contingent knowledge on the related tasks to validate the importance of contingent reasoning. To these ends, we use the Kyoto University Commonsense Inference dataset (KUCI)[1]. KUCI is a Japanese QA dataset with 104k multiple-choice questions that ask contingent relation between basic events directly (Omura et al., 2020). An example is shown in Figure 1. This dataset is also characterized by its semi-automatic data construction method: automatic extraction of contingent pairs of basic event expressions from a web corpus, verification through crowdsourcing, and automatic generation of commonsense inference problems.

It is shown there is a performance gap between humans and computers on this task (Omura et al., 2020). Furthermore, through qualitative evaluation, it has been confirmed computers sometimes answer contingent relation between quite basic events incorrectly. One straightforward approach to alleviating the above problem is to extend the train-

[1] https://nlp.ist.i.kyoto-u.ac.jp/EN/?KUCI

Proceedings of the 29th International Conference on Computational Linguistics, pages 812–823
October 12–17, 2022.

ing data and increase the coverage. However, it is not practical from a cost perspective to increase the number of training examples manyfold using crowdsourcing.

We attempt to improve the performance by omitting crowdsourcing, a bottleneck in data augmentation, and utilizing pseudo-problems generated automatically from unverified contingent pairs of basic event expressions. As a web corpus is scalable, and all of the procedures except crowdsourcing are automatic, we can generate pseudo-problems at scale. It is expected pseudo-problems complement the lack of coverage though some problems are noisy and might be unanswerable.

The second objective of this study is to investigate the effects of learning contingent knowledge on the related tasks. On the premise that contingent reasoning is essential to NLU, we can expect contingent knowledge probably helps improve the performance on other NLU tasks. While the transferability of major English datasets has been studied (Phang et al., 2018; Sap et al., 2019b; Sakaguchi et al., 2020; Pruksachatkun et al., 2020), there is room to explore this dataset in terms of the task and language. We investigate the generality of contingent knowledge through quantitative evaluation of transfer learning on the related tasks.

In summary, we improve commonsense contingent reasoning by straightforward data augmentation. We generated 862k pseudo-problems, which is about ten times as large as the training examples in KUCI (83k), and incorporated them into training. Owing to pseudo-problems, a high-performance pre-trained model has achieved near human-level performance on the commonsense contingent reasoning task. We also investigate the transferability of contingent knowledge to the related tasks. Our experiments demonstrate intermediate-task training on KUCI with pseudo-problems positively affects discourse relation analysis, the Japanese Winograd Schema Challenge, and the JCommonsenseQA, which suggests the importance of contingent reasoning[2].

2 Approach

First, we describe our data augmentation approach to improving commonsense contingent reasoning. Our approach is to automatically generate large-scale pseudo-problems based on the construction method of the Kyoto University Commonsense Inference dataset (KUCI).

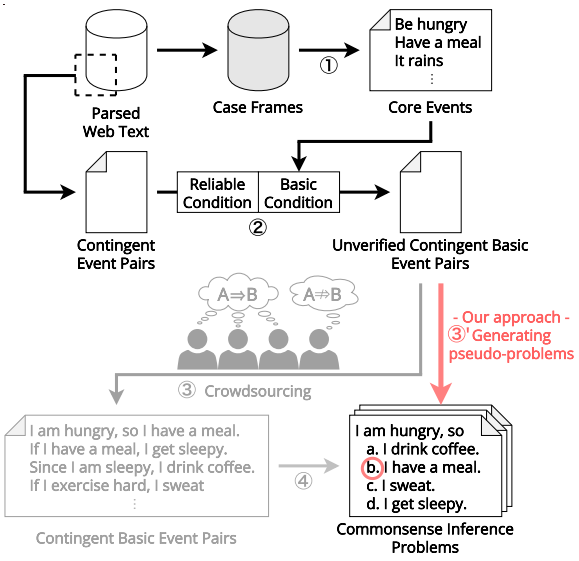

Figure 2: Overview of the method of generating commonsense inference problems in KUCI (gray) and pseudo-problems (red). The further details are described in Omura et al. (2020).

2.1 A Method of Generating Problems

The construction method of KUCI consists of the following four steps (Figure 2).

1. Acquire high-frequency predicate-argument structures (hereafter, **core events**[3]) from *case frames* (Kawahara et al., 2014b).

2. Extract event pairs that are unambiguously connected by explicit discourse markers representing contingent relation and composed of a pair of core events (hereafter, **contingent basic event pairs**).

3. Verify by crowdsourcing whether the extracted event pairs actually have contingent relation or not.

4. Generate problems by taking one of the verified event pairs (hereafter, **base**[3]) and selecting distractors from the latter events of other event pairs that are moderately similar to the base.

In the above procedures, it becomes possible to automatically generate pseudo-problems by omitting step 3 (Figure 2). For the parameters in the

[2]The links to the pseudo-data and code are available at https://nlp.ist.i.kyoto-u.ac.jp/EN/?KUCI

[3]We newly define these terms for clarification.

method, such as the thresholds of frequency for acquiring core events and the conditions on selecting distractors, we set them to the same values as in the construction of KUCI.

2.2 Automatic Extraction of Contingent Basic Event Pairs

We automatically extracted contingent basic event pairs following the method described in Section 2.1. We used a Japanese web corpus containing 3.3 billion sentences as the source text. It had been constructed by crawling web text from 2006 to 2015. There is no overlap of sentences between this corpus and the web corpus used in the construction of KUCI. As a result, we extracted 915k contingent basic event pairs. Omura et al. (2020) reported one-third of the extracted event pairs were removed by crowdsourcing, thus we expect about 600k event pairs to be valid.

2.3 Dealing with Data Leakage

There is a potential issue with generating training data from large-scale text, which is called "Data Contamination" (Brown et al., 2020). This issue is that text may include information about evaluation data, leading to overestimation of model performance.

We deal with this issue by heuristically excluding event pairs that are identical or remarkably similar to the *bases* in evaluation data[4]. Specifically, we apply the following filters based on word order and *core event* pairs.

Filter by word order Exclude an event pair if the length of the overlapping word order between the event pair and any base in evaluation data exceeds 75% of the word count of the base.

Filter by core event pairs Exclude an event pair if the event pair is composed of the core event pair that also composes any base in evaluation data.

For instance, the base of the problem in Figure 1 is "I'm hungry, so → I have a meal at a family restaurant" and composed of the core event pair "be hungry → have a meal at a family restaurant". Let us consider whether the event pair "I'm hungry, so → I have a big meal at the family restaurant" is excluded by the base or not. They have the overlapping word order, {I'm, hungry, so, I, have, a, meal, at, family, restaurant}, of which length (10) exceeds 75% of the word count of the base (11). It is also composed of the same core event pair. Thus, it will be excluded by both filters.

We expect the first filter to exclude syntactically-similar event pairs and the second to exclude those similar in content. As a result of filtering, we acquired 881k contingent basic event pairs.

2.4 Automatic Generation of Pseudo-problems

We went on performing an automatic generation of problems. As a result, we obtained 862k pseudo-problems from the 881k event pairs. The number of the pseudo-problems is about ten times as large as that of the training examples in KUCI (83k).

To analyze the quality of pseudo-problems, we randomly sampled 50 problems and manually evaluated them. As a result of manual evaluation, 36 of 50 problems were judged as answerable, which appears to be sufficient quality for pseudo-data.

3 Experiments

We conducted experiments to investigate the effects of incorporating pseudo-problems into training on the commonsense contingent reasoning task and the related tasks.

3.1 Model

We evaluated the performance of the BERT (Devlin et al., 2019) and XLM-RoBERTa (XLM-R) (Conneau et al., 2020) models.

BERT We employed the NICT BERT Japanese Pre-trained model (with BPE)[5]. It was pre-trained on the full text of Japanese Wikipedia for 1.1 million steps with a batch size of 4,096, partly referring to the pre-training configuration of RoBERTa (Liu et al., 2019). The model architecture is the same as the $\mathrm{BERT}_{\mathrm{BASE}}$.

XLM-R We adopted the XLM-RoBERTa$_{\mathrm{LARGE}}$ model[6], which was pre-trained on a huge multilingual corpus consisting of Wikipedia and CC-100 (Wenzek et al., 2020). The model architecture is the same as the $\mathrm{BERT}_{\mathrm{LARGE}}$, but the embedding layer is relatively large due to its multilingual vocabulary. It is one of the high-performance pre-trained models for Japanese among those publicly available.

[4] To be specific, "evaluation data" refers to the development and test splits of KUCI.

[5] https://alaginrc.nict.go.jp/nict-bert/index.html (in Japanese)

[6] https://huggingface.co/xlm-roberta-large

3.2 Experimental Settings

The hyper-parameters used in the experiments are included in Appendix A.

3.2.1 Commonsense Contingent Reasoning

As is mentioned in Section 1, we used KUCI for assessing commonsense contingent reasoning ability. The task is to select the most appropriate sentence following the context from 4 choices like Figure 1. The dataset contains 83,127 / 10,228 / 10,291 examples for training, development, and test split, respectively.

During the fine-tuning phase, we minimize cross-entropy loss between the scores of each choice normalized by the softmax function and a one-hot vector representing the correct answer as 1. The scores of each choice are computed by inputting pairs of a context and the choice separated by special tokens and converting the hidden representations of the first token ([CLS]) into scalars by a linear transformation. When incorporating pseudo-problems into training, we define the objective function L as the weighted sum of cross-entropy losses of commonsense inference problems and pseudo-problems. The above can be expressed by the following equations.

$$H = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp(\mathbf{s}_{kj})}{\sum_{i=1}^{4}\exp(\mathbf{s}_{ki})}$$
$$L = H_{ci} + \lambda \times H_{pseudo}$$

where N is a batch size, j is the index of a correct choice among 1 to 4, s_{ki} is the score of the i-th choice of k-th example, H is the cross-entropy loss of commonsense inference problems or pseudo-problems, and λ is the weight for pseudo-problems.

During the inference phase, the choice with the highest score is selected as an answer. We evaluated the models by accuracy.

Comparative Method To investigate the effectiveness of a multiple-choice format, we also performed additional pre-training referring to Task-Adaptive Pre-Training (Gururangan et al., 2020). Specifically, we ran an additional Masked Language Modeling (MLM) task on the 881k event pairs used for generating pseudo-problems. For convenience, we name it "AMLM". After the additional pre-training, we fine-tuned the models on KUCI and the related tasks.

3.2.2 Intermediate-Task Transfer Learning

We performed transfer learning from the models fine-tuned on KUCI with pseudo-problems to investigate the effects of learning contingent knowledge. In this study, we employed discourse relation analysis, the Japanese Winograd Schema Challenge (JWSC) (Shibata et al., 2015), and the Japanese CommonsenseQA (JCQA) (Kurihara et al., 2022) as the related tasks.

Discourse Relation Analysis We used the Kyoto University Web Document Leads Corpus (KWDLC)[7] (Kawahara et al., 2014a; Kishimoto et al., 2018) for this task. KWDLC has been built by collecting the first three sentences of various kinds of web documents, and its size amounts to 6,445 documents. All the documents have been annotated with discourse relations between clauses using crowdsourcing. Moreover, 500 of 6,445 documents have also been annotated by linguistic experts. In this study, we used about 37k clause pairs with crowdsourced labels as training data and evaluated the classification performance on 2,320 clause pairs with expert labels.

The task is a seven-way classification of discourse relations between clauses, including "No Relation". We fine-tuned the models following the sentence pair classification framework proposed by Devlin et al. (2019) and ran five-fold cross validation. We used micro-averaged precision, recall, and F1 score computed without examples with the "No Relation" label as evaluation metrics.

JWSC The Winograd Schema Challenge (WSC) is the task to select the antecedent of a pronoun from two candidates (Levesque, 2011). The task itself is coreference resolution but designed to require commonsense reasoning. JWSC[8] is constructed by translating the Rahman and Ng (2012) version of WSC into Japanese.

As we excluded the event pairs containing demonstrative pronouns so as not to generate problems that require more context, there is concern that intermediate-task training on KUCI with pseudo-problems might hurt performance on JWSC due to forgetting the knowledge about demonstratives. Accordingly, we recast JWSC as binary question answering by replacing a pronoun with each antecedent candidate. The resulting dataset is bal-

[7] https://github.com/ku-nlp/KWDLC

[8] https://github.com/ku-nlp/Winograd-Schema-Challenge-Ja

Model	Setting	Acc.
BERT	KUCI	79.3 ± 0.2
	KUCI + Pseudo-problems ($\lambda = 0.1$)	84.1 ± 0.1
	KUCI + Pseudo-problems ($\lambda = 0.5$)	$\mathbf{84.7 \pm 0.1}$
	KUCI + Pseudo-problems ($\lambda = 1.0$)	84.6 ± 0.2
	AMLM → KUCI	83.9 ± 0.1
XLM-R	KUCI	86.0 ± 0.1
	KUCI + Pseudo-problems ($\lambda = 0.1$)	88.5 ± 0.1
	KUCI + Pseudo-problems ($\lambda = 0.5$)	$\mathbf{88.8 \pm 0.1}$
	KUCI + Pseudo-problems ($\lambda = 1.0$)	88.6 ± 0.1
	AMLM → KUCI	86.2 ± 0.2
Human (Omura et al., 2020)		88.9

Table 1: Accuracy on the test split of KUCI. The scores are the mean and standard deviation over three runs with different random seeds. Arrows denote multi-stage fine-tuning. For instance, "AMLM → KUCI" means fine-tuning on KUCI after additional pre-training.

anced and consists of 2,644 / 1,128 examples for training and test split, respectively. Since the development split is not provided, we carried out five-fold cross validation by splitting the training set into 8:2. We trained bert-based logistic regression models and evaluated them by accuracy and Area Under the ROC Curve (AUC).

JCQA JCQA[9] is the Japanese version of CommonsenseQA (Talmor et al., 2019) and consists of 11k five-choice questions regarding a wide range of relations between basic concepts. The questions are based on subgraphs extracted from ConceptNet (Speer et al., 2017) and manually created using crowdsourcing.

Since the task is multiple-choice question answering, we fine-tuned models following the same method described in 3.2.1. We also evaluated the models by accuracy.

3.3 Experimental Results

Commonsense Contingent Reasoning Table 1 shows the experimental results of the commonsense contingent reasoning task. Owing to pseudo-problems, both the BERT and XLM-R models improved the accuracy by 5.4 and 2.8 points, respectively. Notably, the XLM-R model has achieved performance comparable to humans. Putting moderately low weight on pseudo-problems makes the performance slightly better.

Figure 3 shows the learning curves of the models on the development split of KUCI. The crosses

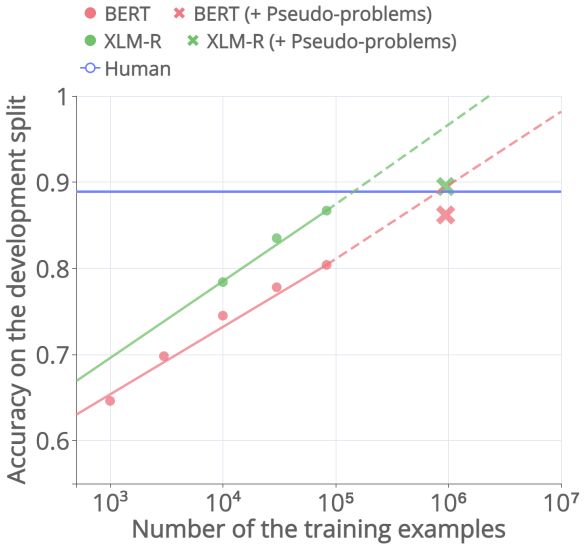

Figure 3: Learning curves of the BERT and XLM-R models on the development split of KUCI. We excluded the degenerate results of the XLM-R model when fine-tuned on a small number of training examples ($N \in \{10^3, 3 \times 10^3\}$).

representing the accuracy on the "KUCI + Pseudo-problems" setting are under the extrapolated learning curves, which implies the difference in quality between the training examples in KUCI and pseudo-problems.

Discourse Relation Analysis As for the related tasks, we can see from Table 2 that intermediate-task training on KUCI with pseudo-problems is effective in discourse relation analysis, particularly in BERT. Since the problems are based on con-

[9] https://github.com/yahoojapan/JGLUE

Model	Setting	Prec.	Rec.	F1
BERT	KWDLC	55.2 ± 2.9	38.4 ± 1.0	45.1 ± 1.1
	KUCI → KWDLC	$\mathbf{58.1 \pm 2.4}$	38.3 ± 1.3	45.7 ± 0.8
	KUCI + Pseudo-problems ($\lambda = 0.5$) → KWDLC	55.9 ± 1.1	$\mathbf{41.0 \pm 2.9}$	$\mathbf{47.0 \pm 2.4}$
	AMLM → KUCI → KWDLC	51.8 ± 3.7	38.4 ± 1.3	43.7 ± 0.7
XLM-R	KWDLC	57.4 ± 1.7	45.5 ± 2.8	50.3 ± 1.3
	KUCI → KWDLC	$\mathbf{57.8 \pm 2.3}$	$\mathbf{48.2 \pm 0.3}$	$\mathbf{51.9 \pm 0.2}$
	KUCI + Pseudo-problems ($\lambda = 0.5$) → KWDLC	57.2 ± 1.0	47.4 ± 1.8	51.5 ± 0.7
	AMLM → KUCI → KWDLC	55.2 ± 1.6	34.5 ± 0.6	40.9 ± 1.0
Human (Crowdworker) (Kishimoto et al., 2020)		54.7	48.6	51.5

Table 2: Performance of discourse relation analysis on KWDLC. The scores are the mean and standard deviation over three runs of five-fold cross-validation with different random seeds. As with Table 1, arrows denote multi-stage fine-tuning. Note that we performed additional Masked Language Modeling (AMLM) on the 881k event pairs used for generating pseudo-problems, not the training examples in KWDLC, to compare the methods of utilizing pseudo-data.

Model	Setting	Ca./Re.	Cond.	Purp.	Justif.	Cont.	Conc.	F1
BERT (ensemble)	KWDLC	76 / 138	32 / 43	18 / 37	0 / 6	2 / 19	54 / 84	46.7
	KUCI → KWDLC	81 / 132	32 / 43	18 / 31	1 / 6	2 / 17	47 / 72	48.0
	KUCI + Pseudo-problems → KWDLC	81 / 139	33 / 49	17 / 29	0 / 4	1 / 12	56 / 85	**48.8**
XLM-R (ensemble)	KWDLC	98 / 159	33 / 46	16 / 34	2 / 4	0 / 18	60 / 88	52.1
	KUCI → KWDLC	109 / 201	34 / 53	18 / 32	3 / 7	0 / 26	56 / 85	51.3
	KUCI + Pseudo-problems → KWDLC	99 / 168	33 / 50	18 / 28	1 / 2	0 / 22	64 / 98	**52.4**
Human (Crowdworker) (Kishimoto et al., 2020)		100 / 175	37 / 54	19 / 44	6 / 32	4 / 30	54 / 67	51.5
Total number of true positives and false negatives		242	54	36	15	6	100	—

Table 3: Detailed results of discourse relation analysis by the ensemble models. The third to eighth columns stand for the discourse relations, "Cause or Reason", "Condition", "Purpose", "Justification", "Contrast", and "Concession", respectively. The values on the left side are the numbers of true positives for the discourse relation, and those on the right side are total numbers of true positives and false positives.

tingent basic event pairs, which are connected by explicit discourse markers representing causal or conditional relation[10], we presume the knowledge about these discourse relations is successfully transferred.

We also describe the detailed results of discourse relation analysis in Table 3. The models transferred from KUCI with pseudo-problems perform better on classifying causal and purpose relations. Compared with crowdworkers, there is room for improvement in precision of concession and infrequent relations.

JWSC The experimental results of JWSC are shown in Table 4. We observed a few degenerate runs[11] (Phang et al., 2018; Pruksachatkun et al., 2020) on the "JWSC" setting despite fine-tuning for 50 epochs. This phenomenon often occurs when training large models on a small dataset, and several studies have reported intermediate-task training can alleviate it (Phang et al., 2018; Pruksachatkun et al., 2020). We also confirmed the same result in this experiment.

We found KUCI is beneficial to JWSC, but pseudo-problems are not necessarily. JWSC contains a non-negligible number of questions regarding concession relation (e.g. "James asked Robert a favor. However, James/Robert declined."), thus we consider putting much emphasis on contingent relation would rather worsen performance. Learning various discourse relations is a promising solution,

[10]These discourse relations are corresponding to "CONTINGENCY:Cause" and "CONTINGENCY:Condition" in the Penn Discourse Treebank (Prasad et al., 2008) and automatically analyzed by the Japanese parser, KNP (Kurohashi and Nagao, 1994).

[11]The training runs that models result in around chance performance. Specifically, we regard less than 0.55 accuracy or AUC as the degenerate runs.

Model	Setting	Acc.	AUC
BERT	JWSC	$66.0 \pm 3.4^{\dagger}$ (68.4 ± 0.1)	$71.4 \pm 4.5^{\dagger}$ (74.5 ± 0.1)
	KUCI → JWSC	**69.9 ± 0.3**	**77.0 ± 0.6**
	KUCI + Pseudo-problems ($\lambda = 0.5$) → JWSC	68.8 ± 1.1	75.0 ± 2.0
	AMLM → KUCI → JWSC	58.1 ± 1.0	61.9 ± 1.1
XLM-R	JWSC	$78.7 \pm 3.2^{\dagger}$ (80.7 ± 0.4)	$85.6 \pm 4.0^{\dagger}$ (88.0 ± 0.5)
	KUCI → JWSC	**81.2 ± 0.1**	**88.7 ± 0.2**
	KUCI + Pseudo-problems ($\lambda = 0.5$) → JWSC	80.0 ± 0.2	**88.7 ± 0.0**
	AMLM → KUCI → JWSC	50.8 ± 0.5	51.7 ± 0.8

Table 4: Accuracy and AUC on the test split of JWSC. The scores are the mean and standard deviation over three runs of five-fold cross-validation with different random seeds. † denotes the results include a few degenerate runs. We also report the results excluding the degenerate runs in parentheses for reference. As for the "AMLM → KUCI → JWSC" setting of XLM-R, the models failed to learn.

which we leave for future work.

JCQA Referring to Table 5, we can see performance gain regarding XLM-R. We speculate it is thanks to the domain match between pseudo-problems and JCQA, considering the report by Kurihara et al. (2022) that pre-training on CC-100 is more effective in JCQA than Wikipedia. Pseudo-problems alone are somewhat insufficient for adapting to the web domain, but they complement some knowledge.

Comparison to AMLM Although AMLM is somewhat effective in KUCI, it is poor at transferring the knowledge[12]. It can be inferred the models learn task-specific knowledge.

3.4 Qualitative Analysis

Figure 4 shows the example problems that BERT got to answer correctly by incorporating pseudo-problems into training. We can see the improvement in accuracy of the problems regarding quite basic contingent relation like Figure 4. The model sometimes gave low scores to all the choices and appeared to choose by elimination, but we observed it became less frequent. We speculate pseudo-problems complement the lack of coverage of the training examples in KUCI. For further information, we include the confusion matrix in Table 6. The improvement is greater though the model got to make a wrong prediction to some problems.

[12]We also tried the "AMLM → related task" setting, but the performance is generally worse than those on the "AMLM → KUCI → related task" setting.

4 Related Work

Owing to large-scale pre-training, the pre-trained models have achieved unprecedented performance on a variety of NLU tasks, including commonsense reasoning (Wang et al., 2019). Besides such improvement in general language understanding, there have been many approaches to improving the performance on commonsense reasoning tasks.

One group of approaches is to utilize automatically created data, to which our approach belongs. For instance, Ye et al. (2019) performed additional pre-training on 16 million fill-in-the-blank multiple-choice questions generated from Wikipedia and ConceptNet (Speer et al., 2017). They improved the performance on two benchmarks for entity-level commonsense reasoning, CommonsenseQA (Talmor et al., 2019) and Winograd Schema Challenge (WSC) (Levesque, 2011), though their method requires the manually constructed resource (ConceptNet). Staliunaite et al. (2021) proposed a data augmentation method for the Choice of Plausible Alternatives (COPA) and its extension (Roemmele et al., 2011; Kavumba et al., 2019), which consists of roughly three steps: filtering web text by several conditions, extracting causal pairs of clauses with the clue of discourse connectives, and generating distractors using language models. They have not investigated the application to the related tasks, focusing on improving commonsense causal reasoning. Shen et al. (2021) improved unsupervised pronoun resolution and commonsense reasoning by pre-training on

Model	Setting	Acc.
BERT	JCQA	81.8 ± 0.1 (82.3)
	KUCI → JCQA	**82.0 ± 0.3**
	KUCI + Pseudo-problems ($\lambda = 0.5$) → JCQA	81.9 ± 0.2
	AMLM → KUCI → JCQA	68.1 ± 0.4
XLM-R	JCQA	84.0 ± 0.5 (84.0)
	KUCI → JCQA	85.0 ± 0.4
	KUCI + Pseudo-problems ($\lambda = 0.5$) → JCQA	**85.3 ± 0.6**
	AMLM → KUCI → JCQA	75.2 ± 0.5
Human (Kurihara et al., 2022)		98.6

Table 5: Accuracy on the **development** split of JCQA. The scores are the mean and standard deviation over three runs with different random seeds. We also include the reported values in the original paper (the numbers in the parentheses) for reference.

霧が晴れると、 (When a fog clears,) **a. 景色が素晴らしい** **(the scenery is amazing)** b. 川の音がすごい (the sound of river is loud) c. 雪遊びも楽しそうだ (playing in the snow sounds nice) d. 写真写りがいまいちだ (it's not photogenic)	嫌な夢を見ると、 (If I have a bad dream,) a. とりあえず寝る (I'll go to bed for now) b. もう寝ます (I'm going to go to bed now) c. さっさと寝ることにする (I'll go to bed quickly) **d. 目を覚まします** **(I'll wake up)**	午後から病院へいくので (I'm going to see a doctor this afternoon, so) a. 滅多に病院に行かない (I rarely see a doctor) b. 土日は勉強に勤しみます (I'll study hard on weekends) **c. 今日は休暇をとる** **(I take a vacation today)** d. 火曜日は眠い (I'm sleepy on Tuesday)

Figure 4: Example problems that the BERT model got to answer correctly by incorporating pseudo-problems into training. The correct choice is bolded, and the choice that BERT previously selected is highlighted in red.

		KUCI	
		correct	incorrect
KUCI + Pseudo-problems ($\lambda = 0.5$)	correct	7,891	1,028
	incorrect	401	908

Table 6: Confusion matrix organizing the numbers of correct and incorrect answers on the development split of KUCI. The matrix shows the results of the BERT model (ensemble).

auto-generated examples that imitate the task.

As for the second objective of this study, there are several studies about the transferability of commonsense knowledge from existing language resources. For instance, it has been reported intermediate-task training on two benchmarks for commonsense reasoning, Social IQA (Sap et al., 2019b) and WinoGrande (Sakaguchi et al., 2020), helps improve the performance on WSC and COPA. Pruksachatkun et al. (2020) showed the datasets that require complex commonsense reasoning, such as CosmosQA (Huang et al., 2019) and HellaSwag (Zellers et al., 2019), are beneficial to several target tasks. Lourie et al. (2021) ran multi-task learning on multiple resources for commonsense reasoning to examine their interactions. Since they have used the datasets that require complex reasoning, they have not focused on a specific type of commonsense reasoning. We focus on commonsense contingent reasoning and investigate the transferability in the language other than English.

5 Conclusion

In this study, we improved commonsense contingent reasoning by incorporating large-scale pseudo-problems into training. We automatically generated 862k pseudo-problems from a Japanese web corpus of 3.3 billion sentences using the existing data

construction method with modification. Owing to pseudo-problems, a high-performance pre-trained model has achieved near human-level performance on the commonsense contingent reasoning task.

We also investigated the effects of learning contingent knowledge on the related tasks: discourse relation analysis, the Japanese Winograd Schema Challenge, and the JCommonsenseQA. Our experiments demonstrated intermediate-task training on KUCI with pseudo-problems has a positive impact on the related tasks, which indicates the importance of contingent reasoning.

We will further analyze what kind of problems current models still answer incorrectly. From the qualitative analysis, we consider building a language resource for evaluating deeper language understanding. As another research direction, it is also tempting to pursue the improvement in NLU by learning various discourse relations between entities or events in documents.

Acknowledgements

We thank anonymous reviewers for their valuable comments. This work was supported by the Japan Kanji Aptitude Testing Foundation. This work was also supported by Information/AI/Data Science Doctoral Fellowship of Kyoto University and Grant-in-Aid for JSPS Fellows #22J15958.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Proc. NeurIPS2020*, volume 33, pages 1877–1901. Curran Associates, Inc.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proc. ACL2020*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proc. ACL2020*, pages 8342–8360, Online. Association for Computational Linguistics.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proc. EMNLP2019*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *Proc. AAAI2021*, pages 6384–6392. AAAI Press.

Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. 2019. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proc. COIN*, pages 33–42, Hong Kong, China. Association for Computational Linguistics.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014a. Rapid Development of a Corpus with Discourse Annotations using Two-stage Crowdsourcing. In *Proc. COLING2014*, pages 269–278, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. 2014b. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In *Proc. EACL2014*, pages 58–67, Gothenburg, Sweden. Association for Computational Linguistics.

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. 2018. Improving Crowdsourcing-Based Annotation of Japanese Discourse Relations. In *Proc. LREC2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yudai Kishimoto, Murawaki Yugo, Daisuke Kawahara, and Sadao Kurohashi. 2020. Japanese Discourse Relation Analysis: Task Definition, Connective Detection, and Corpus Annotation. *Journal of Natural Language Processing*, 27(4):889–931. (in Japanese).

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese General Language Understanding Evaluation. In *Proc. LREC2022*, Marseille, France. European Language Resources Association (ELRA).

Sadao Kurohashi and Makoto Nagao. 1994. KN Parser: Japanese Dependency/Case Structure Analyzer. In *Proceedings of the Workshop on Sharable Natural Language*, pages 48–55.

Hector J. Levesque. 2011. The Winograd Schema Challenge. In *Proc. AAAI2011 Spring Symposium*. AAAI.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In *Proc. AAAI2021*, pages 13480–13488. AAAI Press.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *Proc. ICLR2021*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proc. NAACL2016*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. 2020. A Method for Building a Commonsense Inference Dataset based on Basic Events. In *Proc. EMNLP2020*, pages 2450–2460, Online. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2018. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *CoRR*, abs/1811.01088.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proc. LREC2008*, Marrakech, Morocco. European Language Resources Association (ELRA).

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In *Proc. ACL2020*, pages 5231–5247, Online. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proc. EMNLP2012*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. 2011. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Proc. AAAI2011 Spring Symposium*. AAAI.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *Proc. AAAI2020*, pages 8732–8740. AAAI Press.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *Proc. AAAI2019*, pages 3027–3035. AAAI Press.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense Reasoning about Social Interactions. In *Proc. EMNLP2019*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Ming Shen, Pratyay Banerjee, and Chitta Baral. 2021. Unsupervised Pronoun Resolution via Masked Noun-Phrase Prediction. In *Proc. ACL2021*, pages 932–941, Online. Association for Computational Linguistics.

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2015. Construction and Analysis of Japanese Winograd Schema Challenge. In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, Kyoto, Japan. (in Japanese).

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proc. AAAI2017*, pages 4444–4451. AAAI Press.

Ieva Staliunaite, Philip John Gorinski, and Ignacio Iacobacci. 2021. Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation. In *Proc. AAAI2021*, pages 13834–13842. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *Proc. NAACL2019*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. In *Proc. NeurIPS2021 Datasets and Benchmarks Track (Round 1)*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Proc. NeurIPS2019*, pages 3266–3280. Curran Associates, Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proc. LREC2020*, pages 4003–4012, Marseille, France. European Language Resources Association.

Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. 2019. Align, Mask and Select: A Simple Method for Incorporating Commonsense Knowledge into Language Representation Models. *CoRR*, abs/1908.06725.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proc. EMNLP2018*, 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 *Proc. ACL2019*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

A Hyper-parameters

Table 7, 8, 9, 10, and 11 show the hyper-parameters used in the experiments. We found lower learning rate makes the training of the XLM-R model more stable, thus we set the learning rate of the XLM-R model lower than that of BERT.

Name	Value BERT	Value XLM-R
Epoch		3
Batch size		32
Max sequence length		128
Optimizer		AdamW
Learning rate	2e-5	5e-6
Weight decay	0.01	0.1
Adam's betas params	(0.9, 0.999)	(0.9, 0.98)
Scheduler		Linear decay with linear warmup
Warmup proportion		0.1
Seed		{0, 1, 2}

Table 7: Hyper-parameters for fine-tuning on KUCI and pseudo-problems.

Name	Value BERT	Value XLM-R
Epoch		100
Batch size		256
Max sequence length		128
Optimizer		AdamW
Learning rate		1e-4
Weight decay		0.01
Adam's betas params	(0.9, 0.999)	(0.9, 0.98)
Scheduler		Linear decay with linear warmup
Warmup proportion		0.06
gradient clipping value	-	0.25
Seed		0

Table 8: Hyper-parameters for AMLM. Almost all of the hyper-parameters are referred to Gururangan et al. (2020).

Name	Value BERT	Value XLM-R
Epoch		10
Patience for early stopping		3
Batch size		32
Max sequence length		128
Optimizer		AdamW
Learning rate	2e-5	5e-6
Weight decay	0.01	0.1
Adam's betas params	(0.9, 0.999)	(0.9, 0.98)
Scheduler		Linear decay with linear warmup
Warmup proportion		0.1
Seed		{0, 1, 2}

Table 9: Hyper-parameters for fine-tuning on KWDLC.

Name	Value	
	BERT	XLM-R
Epoch		50
Batch size		32
Max sequence length		128
Optimizer		AdamW
Learning rate	2e-5	5e-6
Weight decay	0.01	0.1
Adam's betas params	(0.9, 0.999)	(0.9, 0.98)
Scheduler		Linear decay with linear warmup
Warmup proportion		0.1
Seed		{0, 1, 2}

Table 10: Hyper-parameters for fine-tuning on JWSC. We set the number of epochs to a large value referring to Mosbach et al. (2021).

Name	Value	
	BERT	XLM-R
Epoch		4
Batch size		32
Max sequence length		128
Optimizer		AdamW
Learning rate		2e-5
Weight decay	0.01	0.1
Adam's betas params	(0.9, 0.999)	(0.9, 0.98)
Scheduler		Linear decay with linear warmup
Warmup proportion		0.1
Seed		{0, 1, 2}

Table 11: Hyper-parameters for fine-tuning on JCQA.