Studies on Data-Driven Discourse Relation Recognition toward Natural Language Understanding

Kazumasa Omura

February 2024

Abstract

The realization of computers that understand natural language to the same extent as humans is the ultimate goal of Natural Language Processing (NLP). Toward this goal, numerous studies strived to build data to train/evaluate some linguistic capability deemed necessary for natural language understanding (NLU) and improve models based on the data.

As text is produced considering the context, it is essential to understand not only the meanings of individual linguistic units (e.g., clauses and sentences) but also the relations between them in order to comprehend the overall meaning. Such semantic relations between text spans are called discourse relations. As discourse relation is an important textual property for understanding the overall meaning of text and has broad applicability, there have been studies focused on discourse relations for a long time.

However, automatic recognition of discourse relations is a long-standing and challenging problem as it requires knowledge about our world beyond natural language. Although various linguistic capabilities of computers have significantly improved with the remarkable development of deep learning in recent years, there is still room for improvement in the linguistic capability to infer discourse relations.

Another problem is that the majority of studies on discourse relations have targeted English even though other languages are also non-negligible. The overemphasis on English may lead to neglect of language-specific phenomena, especially in languages that are linguistic-typologically distant from English such as Japanese, and cause disparity between languages. Therefore, it is worth verifying in the non-English language.

Against such a background, this thesis endeavors to improve the linguistic capability to infer discourse relations primarily in Japanese. Furthermore, we challenge its applications to other NLU tasks and human learning in order to verify the usefulness of discourse relations. To these ends, we explore data generation approaches that are feasible in Japanese.

First, we focus on contingency, which is one of the major discourse relations and crucial for our intellectual activities, and build a Japanese dataset for evaluating the linguistic capability to infer basic contingency (hereafter, commonsense contingency reasoning). Most of the English datasets for commonsense contingency reasoning have been manually built or based on manually constructed language resources. However, this straightforward approach requires a substantial cost and lacks scalability. To solve this issue, we propose a method of semi-automatically generating multiple-choice questions that ask basic contingency from text. Specifically, it is summarized as three steps: automatic extraction of pairs of basic event expressions that have contingent relation from a raw corpus, verification through crowdsourcing, and automatic generation of commonsense contingency reasoning problems from the verified pairs. We build the dataset according to the proposed method and verify its usefulness through experiments.

Next, we work on improving model performance utilizing the constructed dataset. In the aforementioned proposed method, it becomes possible to automatically generate pseudo-problems that imitate commonsense contingency reasoning problems by omitting verification through crowdsourcing. We automatically generate large-scale pseudo-problems by utilizing the scalability and attempt to improve commonsense contingency reasoning by this data augmentation. We also investigate the generality of knowledge about basic contingency through quantitative evaluation by performing transfer learning from a commonsense contingency reasoning task to the related tasks.

Then, we expand our focus from contingency to discourse relations and work on improving discourse relation recognition (DRR). Regarding DRR, one of the biggest issues is the paucity of training data for some error-prone discourse relations. To alleviate this issue, we propose a method of generating synthetic data for these error-prone discourse relations from a large language model. Specifically, it is summarized as two steps: extraction of confusing discourse relation pairs based on false negative rate and generation of synthetic data focused on resolving the confusion. We synthesize data for DRR according to the proposed method and verify its effectiveness through experiments.

Finally, we take up the long-standing problem in Japanese education that elementary school students tend to have an aversion to writing compositions and challenge an educational application in order to ameliorate the situation. Considering the importance of contingency reasoning in NLU, the data constructed in the process of our studies is expected to be useful for human learning as well as machine learning. Thus, we design an AI educational game for elementary school students to study Japanese writing utilizing our constructed data. We also develop smartphone and web applications of the game and conduct a user study to evaluate it.

Acknowledgments

大の国語嫌いだった私が自然言語処理という一見不向きな未知の領域に飛び込み，現在学位論文の謝辞を書いていることは奇妙なことでありますが，この背後には周囲の方々の多大なご支援があったことは言うまでもありません．

まず，学部 4 回生の研究室配属から 6 年間ご指導くださりました黒橋禎夫特定教授に深く感謝いたします．黒橋先生には，常に的確なご指摘で研究をブラッシュアップしていただき，研究が行き詰まった時は建設的なご助言で適切な方向に導いていただきました．また，良い点も至らない点も率直に示していただき，自身の成長を強く促していただきました．どこか自信のない私でしたが，その情熱で大きく後押ししてくださったおかげで結果としてここまで至ることができました．今後は学んだことを広く還元できるように精進いたします．

河原達也教授と楠見孝教授には，学位論文の審査を引き受けていただき，貴重なご助言を賜りました．お礼申し上げます．

椹木哲夫名誉教授と楠見先生には，デザイン学大学院連携プログラムの副指導教員を引き受けていただきました．重ねてお礼申し上げます．

言語メディア研究室の方々には日々多大なご支援をいただきました．河原大輔教授（現：早稲田大学）は右も左も分からない私に研究の基礎を懇切丁寧にご指導くださりました．村脇有吾准教授，Chenhui Chu 特定准教授，Fei Cheng 特定助教には定例の進捗報告や日々のプレゼンテーションで綿密なご助言及びご指摘をいただきました．これらのご指導のおかげで研究を遂行することができました．Yin-Jou Huang 特定助教には様々な場面で建設的なご助言をいただき，時に論文の添削等も引き受けていただきました．清丸寛一特定研究員には学部 4 回生の頃からお世話になり，これまでに多大なご助力をいただきました．ほぼ同時期に博士課程に進学した児玉貴志さん，Qianying Liu さん，Haiyue Song さん，植田暢大さん，Zhuoyuan Mao さんには日々刺激を受け，研究を進める大きな原動力でありました．私が博士

課程在学中に同じく博士課程で精励する先輩の Tareq Alkhaldi さん，後輩の清水周一郎さん，Zhengdong Yang さん，Zhen Wan さんの存在も励みになりました．この他にも，論文紹介などに共に当たったり，研究室で夜遅くまで雑談したりした研究室のメンバー全てに感謝いたします．また，秘書の吉利菜帆さん，石田幸美さん，小杉照美さんには事務手続きを一手に引き受けていただきました．感謝申し上げます．

修士課程および博士課程 1 年目の 3 年間は公益財団法人 日本漢字能力検定協会（漢検）との共同研究に携わらせていただき，本学位論文に関わる研究の大半がこの支援を受けました．漢検側のプロジェクトメンバーの皆様は，私の下手な説明にも優しく耳を傾けてくださり，京大側と異なる視点から貴重なご意見，並びに鋭いご指摘をくださりました．共同研究を通じて他の分野の方々と協調する経験を積むことができたことは，今後大いに役立つと確信しています．また，京大側のプロジェクトメンバーである久保圭准教授（現：早稲田大学），岡久太郎助教（現：静岡大学），Frederic Bergeron さんにもプロジェクトを進める上で多大なご助力をいただきました．改めて感謝申し上げます．

博士課程在学中は株式会社 日本経済新聞社の方々に大変お世話になりました．特に，澤紀彦さん，石原祥太郎さん，白井穂乃さんには，私が進める業務に対して日々ご支援をいただきました．また，定例のミーティングでの雑談は良い息抜きとなり，研究生活の精神的な支えでもありました．感謝申し上げます．

加えて，博士課程在学中に研究助成をいただいた京都大学および独立行政法人日本学術振興会に感謝いたします．

最後に，これまで温かく見守ってくれた家族に感謝の念を表し，謝辞の結びとさせていただきます．

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

The language we listen to, speak, read, and write daily is called natural language as it has naturally emerged and developed in our society. Natural language is a means of cognition and communication for us and thus forms the basis of a broad variety of intellectual activities. In particular, (natural language) text has played an indispensable role in expressing, communicating, and recording information as a medium.

With the proliferation of web media, a vast amount of text is now accumulated every day. The text enables us to consider various applications, whereas it has also become increasingly laborious for us to manually find the information we seek. Against such a background, there is a surge in demand for natural language processing (NLP) technology, which explores the use of computers for processing text.

NLP aims for not only text analysis but also a broad range of applications such as dialogue, machine translation, information retrieval, question answering, and summarization systems. Any of them is supposed to have the following linguistic capabilities, for instance:

- to understand what is written and writer's intention from natural language input.

- to refer to knowledge inside and outside a model and reason as necessary.
- to respond appropriately to the input.

Ultimately, the goal of NLP is to make computers understand natural language to the same extent as humans, for such computers assuredly realize the aforementioned applications and thus enrich our lives. Toward building such a model, NLP has developed through the cycle of building data to train/evaluate some linguistic capability deemed necessary for natural language understanding (**NLU**) and improving models based on the data.

What linguistic capabilities are necessary for NLU? As text is produced considering the context, it is essential to understand not only the meanings of individual linguistic units (e.g., clauses and sentences) but also the relations between them in order to comprehend the overall meaning. Let us consider the following example:

(1) My favorite food is gyoza. I always order it at a ramen restaurant.

When reading this text, we can infer that the second sentence elaborates on the first one. Based on this inference, we can understand that the writer intended to express a strong preference for gyoza.

However, how about the following text?

(2) My favorite food is gyoza. You have a cat.

As there is no semantic relation between the two sentences, we cannot understand the writer's intention behind the text. As shown in the above examples, the linguistic capability to infer semantic relations in text is crucial for NLU. Such semantic relations between text spans are called **discourse relations**.

Discourse relation is an important textual property that supports the coherence of text (Beaugrande and Dressler, 1981) and facilitates understanding writer's intention. In addition to its importance in NLU, a number of discourse relations, such as causal, temporal, and comparative relations, have broad applicability. For instance, causal relations extracted from financial text are potentially profitable for market analysis (Izumi and Sakaji, 2019). Recognizing temporal relations between events described in text is applicable to the automatic generation of timelines. Comparative relations extracted from customer reviews help organize the pros and cons of a product or service. For these reasons, there have been studies focused on discourse relations for a long time.

However, automatic recognition of discourse relations is a long-standing and challenging problem as it requires knowledge about our world beyond natural language. Although various linguistic capabilities of computers have significantly improved with the remarkable development of deep learning in recent years, there is still room for improvement in the linguistic capability to infer discourse relations. For instance, after the advent of general-purpose language models such as BERT (Devlin et al., 2019), it has become possible to acquire a considerable amount of linguistic knowledge through pre-training on large-scale raw corpora. These models have achieved near human-level performance in fundamental analyses, whereas it has also been reported that they do not understand basic discourse relations (Wang et al., 2019; Sap et al., 2019; Bhargava and Ng, 2022). More recently, after the advent of large language models (LLMs) such as GPT-3 (Brown et al., 2020), it has become possible to perform a broad variety of NLP tasks from no or a small number of examples (k-shot learning). However, the k-shot performance of LLMs on the task of identifying discourse relations has been demonstrated to be unsatisfactory (Chan et al., 2023). Thus, this thesis aims to improve the linguistic capability to infer discourse relations toward NLU.

In the modern era where deep learning is thriving, the importance of data for training/evaluation is increasingly growing. The majority of such data has been manually constructed. However, this straightforward approach requires a substantial cost and thus hinders verification in languages other than the major ones, which receive active investment. Actually, the majority of studies on discourse relations have targeted English even though other languages are also non-negligible. The overemphasis on English may lead to neglect of language-specific phenomena (Ruder, 2020), especially in languages that are linguistic-typologically distant from English such as Japanese, and cause disparity between languages. In this thesis, we primarily target Japanese and explore data generation approaches that require minimal manual effort.

On the premise that reasoning about discourse relations is crucial for NLU, the data constructed in the process of our studies is expected to be useful for human learning as well as machine learning. It is worth striving for the social return of the fruits of our studies. We take up the long-standing problem in Japanese education that elementary school students tend to have an aversion to writing compositions and challenge an educational application in order to ameliorate the situation. We also demonstrate the usefulness of discourse relations through the implementation of the application.

In summary, this thesis endeavors to improve the linguistic capability to infer discourse relations primarily in Japanese and verify its importance in NLU. We explore data generation approaches that require minimal manual effort so that we can implement them in Japanese. Furthermore, we challenge an educational application utilizing the data constructed in the process of our studies. The following sections summarize the literature on relevant topics and discuss our approach.

1.2 Natural Language Understanding

Natural Language Understanding (NLU) is an NLP subfield that aims to make computers read natural language input correctly. It includes an extensive variety of NLP tasks as illustrated in Figure 1.1. NLU is often contrasted with Natural Language Generation (NLG), which is another NLP subfield that aims to make computers write natural language output correctly.

1.2.1 Discussion about the Definition of "Language Understanding"

Firstly, it is important to discuss the definition of "language understanding", although we do not have an answer for it. Let us provide a few examples. The definition by Bobrow, who developed the NLU system "STUDENT" in the early days of NLP, is as follows:

> "A computer understands a subset of English if it accepts input sentences which are members of this subset, and answers questions based on information contained in the input." (Bobrow, 1964, p. 2)

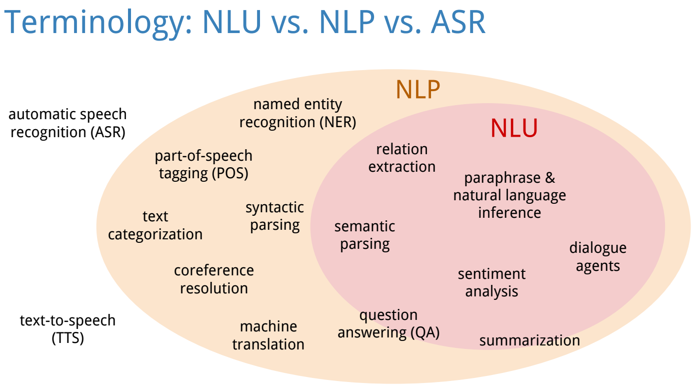


Figure 1.1: Terminology of NLU (MacCartney, 2014, p. 8).

As STUDENT is the rule-based system for solving written algebra questions, the focus at the time was placed on the linguistic capability to respond appropriately to natural language input. The emphasis on responding appropriately to natural language input can also be seen in the core idea of Turing Test (Turing, 1950) that a machine can be considered intelligent if it can convince us through conversation that it is also a human. However, this definition is now debatable as it has been reported that some NLP models return a plausible answer even if they do not fully understand natural language input (Geirhos et al., 2020).

As another example during the development period of NLP, Nagao stated as follows:

> "Language understanding here means to obtain a semantic network from a text." (Nagao, 1997, p. 6)

The semantic network refers to some graphical representation that integrates analysis results of linguistic properties such as dependency, coreference, discourse relation, and so forth. This engineering definition is based on the NLP technology at the time (Sowa, 1992; Bates, 1995) and enables us to quantitatively evaluate language understanding by comparing a predicted semantic network with the correct one. Although it is a constructive suggestion, it is also challenging to establish it in the field.

With the remarkable development of deep learning, which has greatly improved the linguistic capabilities of computers, there is a growing number of attempts to define "language understanding" anew (Bender and Koller, 2020; Bommasani et al., 2021; Merrill et al., 2021). The definition of language understanding by Bender and Koller, who sparked the discussion, is as follows:

> "We take *meaning* to be the relation $M \subseteq E \times I$ which contains pairs (e, i) of natural language expressions e and the communicative intents i they can be used to evoke. Given this definition of meaning, we can now use *understand* to refer to the process of retrieving i given e." (Bender and Koller, 2020, p. 3)

This definition is also paraphrased as "mapping from language to something outside of language". They argued the need for grounding in our world like SHRDLU (Winograd, 1971).

The definition of language understanding has continued to be discussed through ages and varied reflecting generic NLP models at the time. However, it has not yet been well established in NLP.

1.2.2 Bottom-Up Approach to Natural Language Understanding

Instead of defining language understanding, numerous studies have focused on some linguistic capability deemed necessary for NLU and built data to evaluate/improve it. We refer to such an approach (i.e., to work on evaluating/improving some linguistic capability) as *a bottom-up approach.* We introduce Recognizing Textual Entailment and Machine Reading Comprehension as representative examples of NLU tasks that have been addressed with a bottom-up approach. We also mention a recent bottom-up approach to the multifaceted evaluation of NLU.

Recognizing Textual Entailment

Recognizing Textual Entailment (RTE), also known as Natural Language Inference (NLI) (MacCartney, 2009), is the task of identifying entailment and contradiction relations between text spans. It is formulated as a three-way sentence-pair classification of "entailment", "contradiction", or "neutral" class. An example is shown below:

(3) **Premise**: An elderly man selling magazines.
Hypothesis: An old lady selling magazines.
Label: contradiction

In Example (3), the hypothesis is contradictory to the premise as the gender of the subject in each sentence is different; therefore, it is labeled as "contradiction".

Regarding the importance of RTE/NLI in NLU, Condoravdi et al. argued as follows:

> "Relations of entailment and contradiction are the key data of semantics, as traditionally viewed as a branch of linguistics. The ability to recognize such semantic relations is clearly not a sufficient criterion for language understanding: there is more to language understanding than just being able to tell that one sentence follows from another. But we would argue that it is a minimal, necessary criterion." (Condoravdi et al., 2003, p. 1)

Based on such discussion, several RTE/NLI datasets have been built so far (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018; Welleck et al., 2019; Nie et al., 2020). RTE/NLI is still actively studied to evaluate/improve model performance and apply to other NLP tasks, including question answering (Paramasivam and Nirmala, 2022), summarization (Falke et al., 2019), and learning of sentence embeddings (Gao et al., 2021).

Machine Reading Comprehension

Machine Reading Comprehension (MRC) is the task of reading text and answering questions about it. Here is an example retrieved from SQuAD (Rajpurkar et al., 2016, 2018), which is one of the commonly-used MRC datasets:

(4) **Context**: In the past, the Malays used to call the Portuguese Serani from the Arabic Nasrani, but the term now refers to the modern Kristang creoles of Malaysia.
Question: Which term used to refer to Kristang creoles of Malaysia?
Answer: Serani

To answer the above question, it is necessary to recognize terms in the context and choose the most appropriate one according to the question. As shown in the above example, MRC requires an understanding of linguistic properties, which vary depending on questions.

The importance of MRC in NLU has long been acknowledged. For instance, Lehnert stated as follows:

> "Because questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding." (Lehnert, 1978, p. viii)

Indeed, we humans have long tested our understanding through exam questions and studied the effective use of such questions in various fields. Thus, MRC has attracted much attention owing to a shared understanding of its importance, resulting in a prolific NLP task (Zeng et al., 2020).

Recent Bottom-Up Approach to Multifaceted Evaluation of Natural Language Understanding

As these tasks are different components of NLU, combining them is expected to make the evaluation of NLU more robust. Based on such an idea, several studies have constructed a benchmark consisting of a collection of datasets to evaluate NLU from multiple perspectives (Wang et al., 2018, 2019; Srivastava et al., 2023). One of the pioneering and representative examples is the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). GLUE consists of nine existing NLU datasets as organized in Table 1.1 and measures NLU through the average performance on them.

Name	**Task**	**Size**
Single-Sentence Tasks		
CoLA (Warstadt et al., 2019)	linguistic acceptability	11k
SST-2 (Socher et al., 2013)	sentiment analysis	70k
Similarity and Paraphrase Tasks		
MRPC (Dolan and Brockett, 2005)	paraphrase identification	5.8k
STS-B (Cer et al., 2017)	sentence similarity	8.6k
QQP	paraphrase identification	795k
Inference Tasks		
MNLI (Williams et al., 2018)	NLI	432k
QNLI (Rajpurkar et al., 2016)	QA/NLI	116k
RTE (Dagan et al., 2006)	NLI	5.8k
WNLI (Levesque, 2011)	coreference resolution/NLI	0.9k

Table 1.1: Overview of the GLUE benchmark (Wang et al., 2018).

Current NLP models have achieved near human-level performance on the GLUE benchmark, and the target moves on to several subsequent benchmarks consisting of a collection of more challenging and diverse datasets with the development of NLP models. In summary, NLU, and by extension NLP, have developed through the cycle of building data to train/evaluate some linguistic capability deemed necessary for NLU and improving models based on the data.

1.2.3 Our Approach to Natural Language Understanding

In light of this trend, we also take a bottom-up approach to NLU. In order to finally contribute to the multifaced evaluation of NLU, it is important to explore a linguistic capability that is deemed necessary for NLU but relatively underexplored. Therefore, we focus on the linguistic capability to infer discourse relations and work on evaluating/improving it toward NLU.

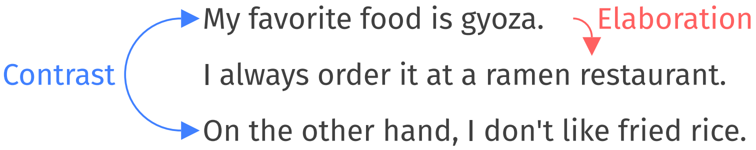

Contrast
My favorite food is gyoza.
Elaboration
I always order it at a ramen restaurant.
On the other hand, I don't like fried rice.

Figure 1.2: Example of discourse relations found in text.

1.3 Discourse Relation

Discourse relation refers to the semantic relation between text spans (cf. Figure 1.2). It is an important textual property for understanding the overall meaning of text. Furthermore, a number of discourse relations are expected to be broadly applicable and actually have been demonstrated to be beneficial to a few NLU tasks and applications (Pan et al., 2018; Saito et al., 2019; Kiyomaru et al., 2020; Tang et al., 2021; Bhargava and Ng, 2022). Nevertheless, the task of identifying discourse relations has not been addressed as actively as RTE/NLI, MRC, and so forth. It is worth focusing on discourse relations toward NLU and other potential applications.

Studies focused on discourse relations have been data-driven, and target discourse relations vary depending on language resources. We introduce existing language resources regarding discourse relations, which are in English unless otherwise noted.

1.3.1 Language Resources regarding Discourse Relations

Several studies have defined discourse relations and built a corpus based on the definitions. Among them, Penn Discourse Treebank and Rhetorical Structure Theory Discourse Treebank are representative corpora.

Penn Discourse Treebank

Penn Discourse Treebank (PDTB) (Prasad et al., 2005, 2008,, 2019) is a corpus built by annotating 2,162 Wall Street Journal (WSJ) articles with discourse relations between adjacent text spans called *arguments*. An example is shown below:

(5) **Arg1**: he'd play for free.
Arg2: You can't give it up that easily,
Relation: Contingency.Cause.Reason
Connective: "because"

Example (5) indicates that the discourse relation between the two arguments is labeled as "Contingency.Cause.Reason" and can be lexicalized by the discourse connective[1] "because". The label "Contingency.Cause.Reason" denotes that the two arguments have three discourse relations of different granularity: "Contingency", "Cause", and "Reason". In summary, PDTB has two major characteristics: discourse relations are defined hierarchically and lexicalized by some discourse connectives. In addition, the two arguments do not contain any discourse connectives, which is called **implicit discourse relation**.

Table 1.2 organizes discourse relations defined in the latest version of PDTB (3.0). Level-1, which is the highest and most coarse-grained, consists of four major classes. Level-2 sub-divides Level-1 into 22 more fine-grained classes, and Level-3 is defined to distinguish Level-2 discourse relations that have directionality (e.g., cause and effect). When using PDTB, the linguistic capability to infer discourse relations is usually measured by the classification performance of Level-1 or Level-2 discourse relations.

Rhetorical Structure Theory Discourse Treebank

Rhetorical Structure Theory Discourse Treebank (RST-DT) (Carlson et al., 2001, 2002) is a corpus built by annotating 385 WSJ articles with discourse structure based on Rhetorical Structure Theory (RST) (MANN and THOMPSON, 1988). Figure 1.3 illustrates an example of discourse annotation based on RST. In RST, text is first segmented into elementary discourse units (EDUs).[2] Next, EDUs are converted into a tree structure by recursively merging adjacent ones. This tree structure is called an RST tree, where each node and edge represent an EDU and

[1] A word or phrase that indicates certain discourse relation such as "and", "but", "for example", and so forth.

[2] EDU refers to the minimal discourse unit, which typically corresponds to a clause.

Level-1	**Level-2**	**Level-3**
TEMPORAL	SYNCHRONOUS	-
	ASYNCHRONOUS	PRECEDENCE SUCCESSION
CONTINGENCY	CAUSE	REASON RESULT NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND ARG2-AS-COND
	CONDITION+SPEECHACT	-
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	-
	PURPOSE	ARG1-AS-GOAL ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	-
	SIMILARITY	-
EXPANSION	CONJUNCTION	-
	DISJUNCTION	-
	EQUIVALENCE	-
	EXCEPTION	ARG1-AS-EXCPT ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST ARG2-AS-SUBST

Table 1.2: Discourse relations defined in PDTB 3.0 (Prasad et al., 2019). For space limitation, we omit the definition of each discourse relation, which can be confirmed in the annotation manual (Webber et al., 2019).

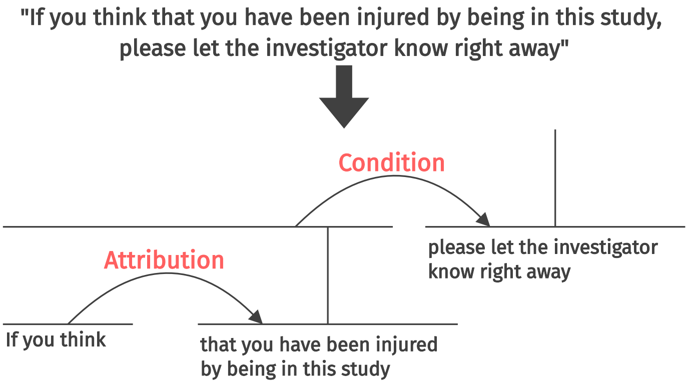

"If you think that you have been injured by being in this study, please let the investigator know right away"
Condition
Attribution
If you think
that you have been injured by being in this study
please let the investigator know right away

Figure 1.3: Example of discourse annotation based on RST.

dependency between EDUs, respectively. Then, the relative importance between nodes of each edge (nuclearity) is analyzed. Finally, discourse relation is assigned to each edge based on its nodes and nuclearity. RST focuses on the discourse structure of text rather than the semantic relations between adjacent text spans (i.e., discourse relations).

Table 1.3 organizes discourse relations defined in RST-DT. In RST-DT, 78 discourse relations are defined and further classified into 16 classes. There are also three syntactic relations defined to impose structure on an RST tree. When using RST-DT, the linguistic capability to infer discourse relations is measured by comparing a predicted RST tree with the correct one.

The major differences between PDTB and RST-DT are summarized as follows:

- PDTB and RST-DT focus on discourse relations and discourse structure of text, respectively.
- PDTB is much larger than RST-DT as PDTB and RST-DT consist of 2,162 and 385 annotated WSJ articles, respectively.
- While PDTB discourse relations are defined hierarchically and lexicalized by some discourse connectives, RST-DT discourse relations consider nuclerity.

PDTB is suitable for extracting local discourse relations. In contrast, RST-DT is fit for applications that require an understanding of the discourse structure of text,

Class	**Relations**
Attribution	attribution, attribution-negative
Background	background, circumstance
Cause	cause, result, consequence
Comparison	comparison, preference, analogy, proportion
Condition	condition, hypothetical, contingency, otherwise
Contrast	contrast, concession, antithesis
Elaboration	elaboration-additional, elaboration-general-specific, elaboration-part-whole, elaboration-process-step, elaboration-object-attribute, elaboration-set-member, example, definition
Enablement	purpose, enablement
Evaluation	evaluation, interpretation, conclusion, comment
Explanation	evidence, explanation-argumentative, reason
Joint	list, disjunction
Manner-Means	manner, means
Topic-Comment	problem-solution, question-answer, statement-response, topic-comment, comment-topic, rhetorical-question
Summary	summary, restatement
Temporal	temporal-before, temporal-after, temporal-same-time, sequence, invertedsequence
Topic Change	topic-shift, topic-drift
Schemata	textual-organization, span, same-unit

Table 1.3: Discourse relations defined in RST-DT (Carlson et al., 2002). For space limitation, we omit the definition of each discourse relation, which can be confirmed in the annotation manual (Carlson and Marcu, 2001).

such as summarization. These and other corpora, such as Discourse Graphbank (Wolf and Gibson, 2005), have been utilized differently according to applications.

Knowledge Bases and Benchmarks regarding Discourse Relations

There are a few knowledge bases and benchmarks regarding discourse relations, which include ASER and DiscoSense. ASER (Activities, States, Events, and their Relations) (Zhang et al., 2020) is a knowledge base built by extracting event pairs that are connected with certain discourse connectives and expanding them using bootstrapping (Agichtein and Gravano, 2000). This knowledge base utilizes discourse connectives that almost uniquely determine their discourse relations referring to PDTB. DiscoSense (Bhargava and Ng, 2022) is a multiple-choice QA dataset consisting of 13k questions that ask the most appropriate sentence following a given context and discourse connective. This dataset measures an understanding of discourse relations through discourse connectives.

Language Resources regarding Discourse Relations in non-English Languages

While both PDTB and RST-DT are in English, similar corpora have also been built in non-English languages with reference to them. Specifically, PDTB-style corpora have been built in Turkish (Zeyrek and Webber, 2008), Hindi (Oza et al., 2009), French (Danlos et al., 2012), Cheze (Poláková et al., 2013), Chinese (Zhou and Xue, 2015; Jiang et al., 2018), and so forth. Their annotation schemes differ, reflecting the characteristics of each language.[3] There also exist RST-style corpora in German (Stede, 2004; Stede and Neumann, 2014), Spanish (da Cunha et al., 2011), Dutch (Vliet et al., 2011), and so forth. As shown in the above examples, studies on discourse relations have developed based on representative corpora.

1.3.2 Language Resources regarding Contingency

Contingency is the discourse relation between events established when one is likely to cause the other. It covers a broad range of causal relations and thus is crucial for our intellectual activities. Owing to its importance and broad applicability, several studies have constructed language resources regarding contingency apart

[3] For instance, Chinese PDTB (Zhou and Xue, 2015) adopts flat 11-way discourse relations and skips annotating discourse connectives of implicit discourse relations.

from the aforementioned ones.

For instance, ATOMIC (Sap et al., 2019) is a commonsense knowledge base comprising 877k pairs of basic event expressions that have contingent relation. This knowledge base has been constructed by extracting frequent event expressions from corpora and manually annotating contingent relations associated with each event by crowdsourcing. The updated version of ATOMIC (Hwang et al., 2021) incorporates a part of ConceptNet (Speer et al., 2017), which accumulates various relations between basic concepts. Thus, it contains a broad range of contingent relations between entities and events as organized in Table 1.4.

There also exist several benchmarks regarding contingency such as COPA, SWAG, and Social IQA. COPA (Choice Of Plausible Alternatives) (Roemmele et al., 2011) is a QA dataset consisting of 1k questions that ask causal relation between daily events. Each question is manually created by experts. SWAG (Situations With Adversarial Generations) (Zellers et al., 2018) is a multiple-choice QA dataset comprising 113k questions that ask the most appropriate verb phrase following a given context. This dataset measures an understanding of contingent relation between actions through inferring a consecutive verb phrase. Social IQA (Social Intelligence QA) (Sap et al., 2019) is also a multiple-choice QA dataset consisting of 38k questions regarding social commonsense knowledge. Each question is manually created by crowdsourcing based on a triple from ATOMIC.

1.3.3 Current Status of Discourse Relation Recognition

The linguistic capability to infer discourse relations has been measured through a discourse relation recognition task. Discourse Relation Recognition (DRR) is the task of identifying the discourse relation given a pair of text spans, which is often evaluated using a PDTB-style corpus. In particular, that focused on implicit discourse relations is called Implicit Discourse Relation Recognition (IDRR). The following paragraphs outline the current status of DRR.

One of the major recent turning points in NLP is the advent of general-purpose language models such as BERT. BERT (Devlin et al., 2019) is a Transformer-based encoder-only model (Vaswani et al., 2017) that receives a sequence of tokens and returns a contextualized embedding of each token. This and subsequent models

	HEAD	Relation	Tail	Size
PHYSICAL-ENTITY	bread	ObjectUse	make french toast	165,590
		AtLocation	basket; pantry	20,221
		MadeUpOf	dough; wheat	3,345
		HasProperty	cooked; nice to eat	5,617
	baker	CapableOf	coat cake with icing	7,968
		Desires	quality ingredients	2,737
		Not Desires	bad yeast	2,838
EVENT-CENTERED	X runs out of steam	IsAfter	X exercises in the gym	22,453
		HasSubEvent	become tired	12,845
		IsBefore	X hits the showers	23,208
		HinderedBY	drinks too much coffee	106,658
		Causes	takes a break	376
		xReason	did not eat breakfast	334
	X watches ___ anyway	isFilledBy	bad yeast	33,266
SOCIAL-INTERACTION	X runs out of steam	xNeed	do something tiring	128,955
		xAttr	old; lazy; lethargic	148,194
		xEffect	drinks some water	115,124
		xReact	tired	81,397
		xWant	to get some energy	135,360
	X votes for Y	xIntent	to give support	72,677
		oEffect	receives praise	80,166
		oReact	grateful; confident	67,236
		oWant	thank X; celebrate	94,548

Table 1.4: Social and physical commonsense relations defined in ATOMIC 2020 (Hwang et al., 2021). They include a number of contingent relations such as "Causes".

have become capable of acquiring a considerable amount of linguistic knowledge through pre-training on large-scale raw corpora. Fine-tuning according to a downstream task has led to new state-of-the-art performance on various NLP tasks.

This improvement was no exception in DRR (Xiang and Wang, 2023). However, it has also been reported that these models are incapable of correctly answering some questions that ask basic discourse relations (Wang et al., 2019; Sap et al., 2019; Bhargava and Ng, 2022).

Another turning point is the advent of large language models (LLMs) such as GPT-3. GPT-3 (Brown et al., 2020) is a Transformer-based decoder-only model that autoregressively generates tokens following a sequence of input tokens. These LLMs have become capable of performing a broad variety of NLP tasks from no or a small number of examples by scaling up the model and data size.

However, the k-shot performance of LLMs in IDRR has been demonstrated to be far behind the fine-tuning performance of much smaller language models. Specifically, the zero-shot performance of ChatGPT on PTDB 2.0 is Micro-F1 of 27.0 (Chan et al., 2023), whereas BERT has achieved Micro-F1 of 51.4 (Kishimoto et al., 2020).

As outlined above, DRR, especially IDRR, is a long-standing and challenging problem. In addition, this is the current status of DRR in English; there is much room for exploration in non-English languages.

1.3.4 Our Approach to Discourse Relation Recognition

Against the above background, we work on improving the linguistic capability to infer discourse relations primarily in Japanese. Considering that the performance of neural language models empirically improves in proportion to the amount of training data (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2020; Henighan et al., 2020; Bahri et al., 2021), it is important to delve into methodology to construct high-quality data. Therefore, we explore data generation approaches that require minimal manual effort so that we can implement them in Japanese.

1.4 Outline of the Thesis

The objectives of this thesis are summarized as two folds: to improve the linguistic capability to infer discourse relations primarily in Japanese and to explore its applications to other NLU tasks and human learning. The rest of this thesis describes our studies toward these objectives.

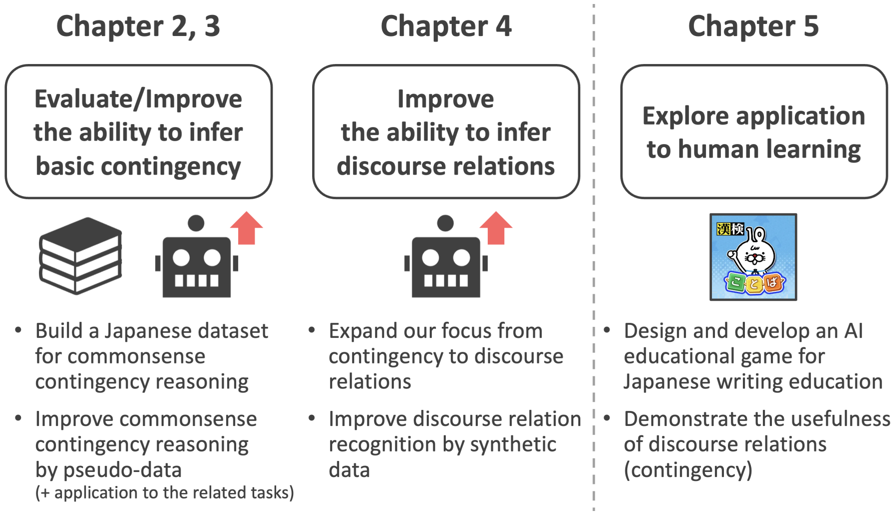

Chapter 2, 3
Evaluate/Improve the ability to infer basic contingency
Build a Japanese dataset for commonsense contingency reasoning
Improve commonsense contingency reasoning by pseudo-data
(+ application to the related tasks)
Chapter 4
Improve the ability to infer discourse relations
Expand our focus from contingency to discourse relations
Improve discourse relation recognition by synthetic data
Chapter 5
Explore application to human learning
Design and develop an AI educational game for Japanese writing education
Demonstrate the usefulness of discourse relations (contingency)
漢検
ことば

Figure 1.4: Outline of the thesis.

In Chapter 2, we present our work on building a Japanese dataset focused on the linguistic capability to infer basic contingency (hereafter, commonsense contingency reasoning). Contingency is the discourse relation between events established when one is likely to cause the other. Despite its importance and broad applicability, there has existed no large-scale Japanese dataset for commonsense contingency reasoning. To solve this issue, we propose a method of semi-automatically generating multiple-choice questions that ask basic contingency from text and build a Japanese dataset according to the proposed method.

In Chapter 3, we work on improving model performance utilizing the constructed dataset. We automatically generate large-scale pseudo-problems by utilizing the scalability of the aforementioned proposed method and attempt to improve commonsense contingency reasoning by this data augmentation. We also investigate the generality of knowledge about basic contingency through quantitative evaluation by performing transfer learning from a commonsense contingency reasoning task to the related tasks.

In Chapter 4, we present our work on improving discourse relation recognition (DRR) by synthetic data. As near human-level performance has been achieved on our constructed dataset thanks to pseudo-problems, we expand our focus from contingency to discourse relations and work on improving DRR. Regarding DRR, one of the biggest issues is the paucity of training data for some error-prone discourse relations. To alleviate this issue, we propose a method of generating synthetic data for these error-prone discourse relations using a large language model.

In Chapter 5, we introduce an educational application utilizing the data constructed in the process of our studies. Considering the importance of contingency reasoning in NLU, the data constructed in the process of our studies is expected to be useful for human learning as well as machine learning. Thus, we take up the long-standing problem in Japanese education that elementary school students tend to have an aversion to writing compositions and challenge an educational application utilizing our constructed data in order to ameliorate the situation.

In Chapter 6, we conclude this thesis and discuss the future prospects of our studies.

Chapter 2

Building a Commonsense Contingency Reasoning Dataset

2.1 Introduction

The realization of natural language understanding (NLU) by computers is the ultimate goal of natural language processing (NLP). Toward this goal, there have been numerous studies that consider task settings to train/evaluate NLU by computers and build the data (Wang et al., 2018, 2019; Srivastava et al., 2023). In such initiatives, it has been argued that acquiring knowledge about both language (e.g., syntax and meanings of words and phrases) and our world beyond language is necessary for the realization of NLU by computers.

After the advent of general-purpose language models such as BERT (Devlin et al., 2019), the problem of acquiring knowledge about language has been solved to a large extent through pre-training on large-scale raw corpora. It is now possible to represent the meaning of a word according to its context as a vector. Fine-tuning based on these vectors has led to near human-level performance in natural language inference (NLI), shallow question answering, and so forth.

On the other hand, there are still some issues left with acquiring knowledge about our world beyond language. As it is open-ended, the data focused on the fundamental part of it (i.e., commonsense knowledge) has been actively constructed. One of the issues there is how to focus on commonsense knowledge in order to acquire it from extensive knowledge.

Several approaches have been attempted to guarantee such generality of knowledge. SWAG (Zellers et al., 2018), for instance, focuses on knowledge about daily events that can be visually perceived by utilizing video captions. However, this approach limits the range of knowledge that can be acquired. CommonsenseQA (Talmor et al., 2019) is based on the basic vocabulary that is covered by ConceptNet (Speer et al., 2017). This approach lacks scalability as it can create only 12k questions from the whole ConceptNet.

Another issue is that biases in dataset construction must be reduced as much as possible. In the above two approaches, distractor sentences or questions were produced from a language model or by crowdsourcing, which induces generation bias of a language model or annotation artifacts[1] (Gururangan et al., 2018).

To solve these issues, we attempt to generate problems from text (a raw corpus) rather than create problems manually or based on manually constructed language resources. We propose a method of extracting pairs of basic event expressions that have contingent relation from a raw corpus, verifying them through crowdsourcing, and generating multiple-choice questions from the verified pairs.

Basic event expressions (hereafter, **basic events**) are defined as events (Saito et al., 2018; Kiyomaru, 2022) composed of high-frequency predicate-argument structures that are extracted from a raw corpus and aggregated by clustering according to their usages. According to this definition, we automatically extract pairs of basic events that have contingent relation with the clue of discourse connectives. We call them **contingent basic event pairs**.

Examples of contingent basic event pairs are shown below.

(1) a. I am hungry, so I have a meal.
    b. If I have a meal, I get sleepy.

[1] Certain patterns (biases) of vocabulary, style, and so forth contained in crowdworkers' writing.

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

Figure 2.1: Example of a commonsense contingency reasoning problem. ✓ denotes the correct choice.

c. Since I am sleepy, I drink coffee.
d. If I exercise hard, I sweat.

Based on these contingent basic event pairs, we can generate a **commonsense contingency reasoning problem** by adopting the latter events of other pairs as distractors (cf. Figure 2.1).

As the proposed method is based on automatic extraction from a raw corpus, it is scalable and does not limit the domain. In addition, there is no bias induced by crowdsourcing because we ask crowdworkers to just verify sentences. Furthermore, the proposed method is relatively language-independent as it does not depend heavily on crowdsourcing or manually constructed language resources, and discourse connectives are ubiquitous in various languages.

In this study, we construct a Japanese commonsense contingency reasoning dataset by applying the proposed method to a Japanese web corpus. We verify its usefulness through experiments.

The contributions of this study are summarized as follows:

- We propose a semi-automatic (scalable and low-cost) method for building a commonsense contingency reasoning dataset, which combines automatic extraction from a raw corpus and crowdsourcing.

- According to the proposed method, we built a Japanese commonsense contingency reasoning dataset comprising 104k multiple-choice questions from a Japanese web corpus.

- We confirmed that there was a reasonable gap in the linguistic capability to infer basic contingency between computers and humans and negligible bias in the constructed dataset.[2]

2.2 Related Work

Existing language resources for commonsense reasoning can be classified into knowledge bases and QA datasets.

2.2.1 Knowledge Bases for Commonsense Reasoning

Commonsense knowledge bases have been constructed by experts, crowdsourcing, or games with a purpose. They include Cyc, ConceptNet, and ATOMIC.

Cyc (OpenCyc) (Lenat, 1995) is a commonsense knowledge base that accumulates commonsense knowledge transcribed by experts in certain notation, the size of which amounts to 2.4 million triples. Although the quality is high owing to manual construction by experts, it takes a considerable amount of time to construct this knowledge base.

ConceptNet (Speer et al., 2017) is a multilingual commonsense knowledge base that is primarily sourced from Open Mind Common Sense (OMCS) (Singh et al., 2002), the size of which amounts to 2.8 million triples with two English concepts (Otani et al., 2018). This knowledge base incorporates OpenCyc and thus includes various relations between basic concepts, although the number of contingent relation between events is not large.

ATOMIC (Sap et al., 2019) is a commonsense knowledge base comprising 877k pairs of basic events that have contingent relation (cf. Section 1.3.2). The authors collected these pairs by crowdsourcing based on frequent events extracted from corpora. This knowledge base has been updated for further coverage (Hwang et al., 2021).

These fully manual or crowdsourcing approaches require a substantial cost and lack scalability. In addition, how to incorporate such knowledge bases into an NLP model has been studied but has not been established yet.

[2] We named the constructed dataset "the Kyoto University Commonsense Inference dataset (KUCI)" and released it to the public (`https://nlp.ist.i.kyoto-u.ac.jp/EN/?KUCI`).

2.2.2 QA Datasets for Commonsense Reasoning

Numerous QA datasets for commonsense reasoning have been built so far. They include COPA, SWAG, and CommonsenseQA.

COPA (Roemmele et al., 2011) consists of 1k two-choice questions that ask causal relation between daily events. Each question presents a premise sentence and requires to choose its cause or effect sentence from two alternatives. This dataset has been manually created for the purpose of evaluation and is too small to learn commonsense knowledge.

SWAG (Zellers et al., 2018) is a commonsense reasoning dataset comprising 113k multiple-choice questions that ask the most appropriate verb phrase following a given context. Questions were created from video captions to ensure the target knowledge is common sense; thus, the domain of the dataset is limited to physical phenomena. Each question is based on two consecutive sentences extracted from video captions, where the first sentence and the subject of the second sentence constitute context, and the rest is regarded as a correct choice. The authors generated distractors from a language model and removed those that were easily discriminated by a discriminative model for quality control. However, SWAG was solved by BERT with near human-level performance. This is attributed to biases that are embedded in distractors by an LSTM-based language model (Hochreiter and Schmidhuber, 1997) and detected by BERT (Zellers et al., 2019). They built HellaSwag (Zellers et al., 2019) anew using a superior language model to make biases undetectable by BERT. However, it has also been reported that HellaSwag also contains similar biases (Tamborrino et al., 2020). The bias issue has not been solved yet.

CommonsenseQA (Talmor et al., 2019) is a commonsense reasoning dataset consisting of 12k multiple-choice questions that ask the most appropriate concept[3], given a question. Each question is manually created by crowdsourcing based on a subgraph extracted from ConceptNet. The subgraph consists of one source concept and three target concepts connected with the same relation. A crowdworker writes a question sentence that contains the source concept and whose answer is only one of the target concepts. This approach depends on the manually con-

[3] Usually, a word or phrase.

structed language resource, ConceptNet, and lacks scalability. Furthermore, it may induce annotator bias in question sentences (Geva et al., 2019) as the load of creating question sentences is heavy for crowdworkers.

In addition to these datasets, there exist others focused on certain kind of commonsense reasoning ability, such as Social IQA (Social Intelligence QA) regarding social commonsense knowledge (Sap et al., 2019) and PIQA (Physical Interaction: Question Answering) regarding physical commonsense knowledge (Bisk et al., 2020). There also exist several datasets that do not directly evaluate commonsense reasoning ability but require commonsense knowledge to answer a question, which include Winograd Schema Challenge (WSC) (Levesque, 2011) and WinoGrande that was built by creating 44k WSC-style questions using crowdsourcing (Sakaguchi et al., 2020). Any of these datasets is manually created or based on manually created language resources, and the aforementioned issues may be pointed out.

While most of the above datasets are multiple-choice QA tasks ranging from two to five choices, there have also been attempts to evaluate commonsense reasoning ability as a generative task. CommonGen (Lin et al., 2020), which is one of the representative examples, consists of 35k constrained sentence generation problems created using image captions, ConceptNet, and crowdsourcing. Specifically, given several words that express objects or actions, the task is to generate sentences that describe everyday situations using all the words. For instance, given the words “dog”, “frisbee”, “catch”, and “throw”, the expected behavior is to generate a sentence like “A dog leaps to catch a thrown frisbee”. A generative task evaluates commonsense reasoning ability more directly, whereas there has been no established automatic evaluation metric yet, which makes it challenging to compare model performance. Furthermore, it is often impractical to prepare several references for each problem in order to make automatic evaluation more robust. A multiple-choice QA task has the advantage of making it easy to objectively compare model performance based on accuracy though it limits output candidates.

2.3 Proposed Method

A commonsense contingency reasoning problem consists of a context and four choices. The task is to choose the most appropriate choice as the continuation of a given context, as illustrated in Figure 2.1. We humans infer basic contingency in everyday situations, such as reading text and having a conversation; therefore, the problem examines the linguistic capability crucial for NLU.

To ensure the target knowledge is common sense to some extent, these problems are based on basic events and generated from contingent basic event pairs verified through crowdsourcing. In addition, we combine automatic extraction from a raw corpus and verification through crowdsourcing to avoid impairing scalability and inducing unintended biases. The proposed method of generating commonsense contingency reasoning problems consists of the following four steps (cf. Figure 2.2):

1. Acquire high-frequency predicate-argument structures (hereafter, **core events**) from case frames (Kawahara and Kurohashi, 2006; Kawahara et al., 2014).

2. Extract contingent basic event pairs, event pairs that are unambiguously connected by some discourse connectives representing contingent relation and composed of core events, from parsed text.

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

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

The following subsections describe the details of each step.

2.3.1 Acquisition of Core Events

Basic events in this study are defined as events composed of high-frequency predicate-argument structures (core events) that are extracted from a raw cor-

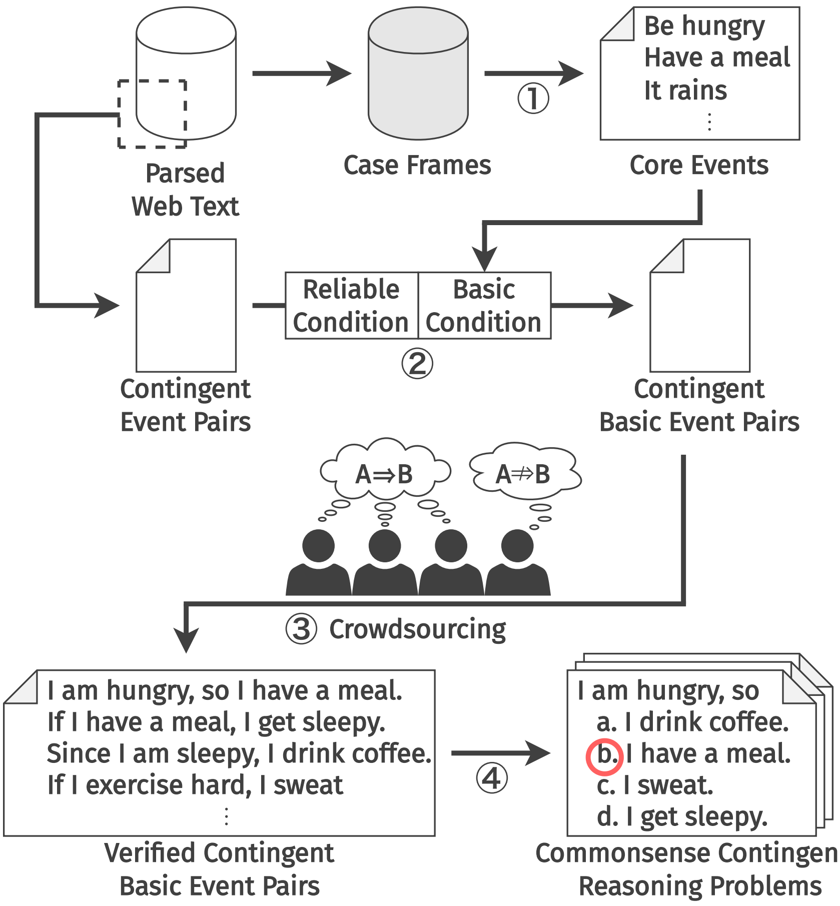


Figure 2.2: Overview of the proposed method of generating commonsense contingency reasoning problems.

pus and aggregated by clustering according to their usages. As the source of core events, we employ case frames (Kawahara and Kurohashi, 2006; Kawahara et al., 2014), which 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 2.1 exemplifies a few case frames of the Japanese verb "壊す [*kowasu*]".

In this study, we extract high-frequency predicate-argument structures from

Case frame	**Case slots**	**Case fillers**
kowasu-1 (injure)	*ga* $_{1756}$ *wo* $_{70135}$ *de* $_{3941}$	I $_{83}$, person $_{65}$, ... stomach $_{25643}$, body $_{17242}$, ... , kidney $_{85}$, ... stress $_{297}$, eating $_{174}$, ...
kowasu-2 (destroy)	*ga* $_{502}$ *no* $_{10147}$ *wo* $_{18274}$	person $_{42}$, Japan $_{42}$, ... place $_{873}$, room $_{851}$, ... atmosphere $_{8140}$, impression $_{3774}$, ...
...		

Table 2.1: Examples of Japanese case frames. *ga*, *wo*, *de*, and *no* roughly correspond to nominative, accusative, instrumental, and genitive cases, respectively. The number following a case or a case filler represents its frequency. For space limitation, examples are expressed only in English.

Case frame	**Case slots**	**Case fillers**
kowasu-1 (injure)	*wo*	stomach, body
kowasu-2 (destroy)	*no* *wo*	place, room atmosphere, impression

Table 2.2: Examples of core events acquired from the case frames in Table 2.1.

case frames as core events. First, top-α frequent predicates in active voice are extracted from the case frame data. For each predicate, case frames, case slots, and case fillers are chosen in decreasing order of frequency until the cumulative sum of frequencies reaches β%, γ%, and δ%, respectively. For instance, case frames are chosen until covering β% of the frequency of a target predicate. These thresholds are empirically set according to a target language.

Table 2.2 shows examples of core events acquired from the case frames in Table 2.1. The parameters for acquiring Japanese core events are described in Section 2.4.1.

2.3.2 Extraction of Contingent Basic Event Pairs

We apply dependency and discourse parsing to a raw corpus and automatically extract event pairs connected with both dependency and contingency. **Event** is the linguistic unit that expresses a single action or state and roughly corresponds to a clause or predicate-argument structure with some modifiers (Saito et al., 2018; Kiyomaru, 2022). The contingent relation between events should be expressed by some discourse connective and causal or conditional relation, corresponding to "Contingency.Cause" or "Contingency.Condition" in Penn Discourse Treebank (Prasad et al., 2008, 2019).

To choose a reliable part from analysis results and extract commonsense event pairs, we filter event pairs by the following conditions. Here, we call the first event in each event pair that represents a cause, reason, or condition **former event** and the second event **latter event**.

Reliable The former and latter events are unambiguously connected.

In the case that only two clauses (events) 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 according to a language-dependent criterion. The criterion for Japanese is described in Section 2.4.2.

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

This condition can be applied in a straightforward manner, 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 attempt to recover it with any of the arguments in the former event and examine whether the recovered latter event is composed of a core event.

For instance, let us consider the event pair "Glass breaks on impact → I replace it". In this case, we generate the recovered latter events "I replace glass" and "I replace impact" by substituting an argument in the former event for "it". We then examine whether either of them is composed of a core event and extract this event pair because "replace glass" is a core event.

Finally, the following post-processing is performed so that crowdworkers in the next step can more accurately judge event pairs.

- Count the frequency of core events contained in (unverified) contingent basic event pairs and exclude event pairs that contain one of the high-frequency core events to remove those that are trivial or contain web-specific functional expressions. For instance, “問題がない (have no problem)” and “情報が満載 (have much information)” are detected as high-frequency trivial core events in Japanese.
- Exclude event pairs that contain demonstratives or unknown words.
- Deduplicate event pairs based on pairs of predicate-argument structure constituting each event pair.

2.3.3 Verification of Contingent Basic Event Pairs through Crowdsourcing

We verify contingent basic event pairs through crowdsourcing. Specifically, we ask crowdworkers to select one of the following two options for each event pair.

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

Here, “A” and “B” denote the former and latter events, respectively.

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

2.3.4 Generation of Commonsense Contingency Reasoning Problems

We automatically generate commonsense contingency reasoning problems from the verified contingent basic event pairs. We take one of the verified pairs (base) and use the former event as a context and the latter event as a correct choice. Distractors are automatically selected from the latter events of other pairs.

In general, remarkably similar distractors to the correct choice are not distinguishable even by humans. On the other hand, dissimilar distractors can be easily distinguished by computers. Thus, we choose distractors that are moderately similar to the correct choice under the following conditions.

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

This similarity is computed 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 the range $\mathrm{RANGE}_{\mathrm{context}}$.

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

To improve the appearance of problems, we choose distractors whose ratio of the number of words against the correct choice is in the range $\mathrm{RANGE}_{\mathrm{length}}$.[4]

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 base.

2.4 Building a Japanese Dataset

We built a Japanese commonsense contingency reasoning dataset according to the proposed method described in Section 2.3.

2.4.1 Acquisition of Core Events

We extracted Japanese core events from the Kyoto University case frames,[5] which had been constructed from 10 billion sentences in web domain. We set the thresh-

[4] As a result of the preliminary experiment, we confirmed that this condition did not affect model performance. Hence, we do not investigate the effect of this condition.

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

olds α, β, γ, and δ to 5,000, 75, 50, and 50, respectively. As a result, we acquired about 14 million core events from 28,642 case frames. Examples of the acquired core events have already been shown in Table 2.2.

2.4.2 Extraction of Contingent Basic Event Pairs

We automatically extracted contingent basic event pairs from a Japanese web corpus comprising 0.7 billion sentences. This corpus is part of an in-house corpus that has been constructed by crawling web text from 2006 to 2015. First, we used the Japanese analyzer, KNP[6] (Kurohashi and Nagao, 1994), and EventGraph[7] (Saito et al., 2018) to extract event pairs from the corpus. KNP performs dependency parsing and labels explicit discourse relations between clauses (events) based on discourse connectives, and EventGraph is the tool that formats analysis results by KNP into event units. As a result, 85 million contingent event pairs were extracted.

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

Finally, we performed the post-processing and extracted 164,910 contingent basic event pairs. The detailed statistics are organized in Table 2.3.

Preliminary Investigation of Basic Condition To investigate the effectiveness of the Basic condition, we randomly selected 100 event pairs from "+Reliable" and "+Reliable+Basic" in Table 2.3, and manually evaluated them. For convenience, we name each set of the selected event pairs "R" and "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 knowledge. Here are examples in "R" excluded by the Basic condition:

(2) 魔力カウンターの乗っていない「魔法都市エンディミオン」に対してサイクロンを発動すると → 破壊できる

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

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

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 2.3: Detailed statistics regarding extraction of contingent basic event pairs. For instance, the value of "+Reliable" represents the number of contingent event pairs that satisfy the Reliable condition.

(activate Cyclone against "Magical Citadel of Endymion" without a spell counter on it→can destroy {it})[8]

(3) すると，→泣きやすい
(do→be quick to tears)

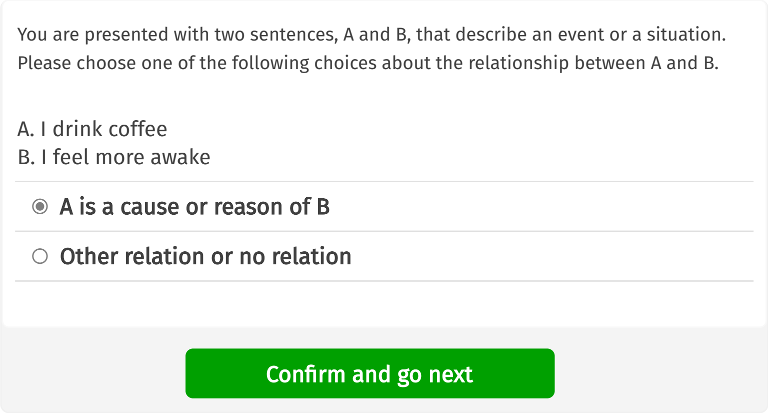
Thanks to the Basic condition, we can remove event pairs that contain domain-specific expressions or lack essential complements. In this sense, the Basic condition is effective in acquiring commonsense knowledge.

2.4.3 Verification of Contingent Basic Event Pairs through Crowdsourcing

We sorted out contingent basic event pairs through crowdsourcing. Regarding a crowdsourcing service, we used Yahoo! Crowdsourcing.[9] Each crowdworker was presented with 17 event pairs per task and chose one from the two options for each event pair (cf. Figure 2.3). 2 out of the 17 event pairs were attention-check with a hidden ground truth, and the answers of crowdworkers who incorrectly judged these event pairs were excluded. Each event pair was verified by four crowdworkers, and we chose the event pairs two or more of whose evaluations are "A is a cause or reason of B".

[8] {} indicates a dropped pronoun.

[9] `https://crowdsourcing.yahoo.co.jp/`



You are presented with two sentences, A and B, that describe an event or a situation. Please choose one of the following choices about the relationship between A and B.

A. I drink coffee
B. I feel more awake

● A is a cause or reason of B

○ Other relation or no relation

Figure 2.3: Crowdsourcing interface for verifying contingent basic event pairs (English translated version).

As a result of crowdsourcing, 104,266 out of 164,910 contingent basic event pairs were chosen, which indicates that approximately one-third of the pairs were removed. This ratio roughly corresponds to the result of the aforementioned preliminary investigation of the Basic condition. The total cost of crowdsourcing was 484,000 JPY, and the cost per problem was 4.7 JPY.

2.4.4 Generation of Commonsense Contingency Reasoning Problems

Finally, we automatically generated commonsense contingency reasoning problems from the verified contingent basic event pairs. The similarity range $\text{RANGE}_{\text{choice}}$ in the condition Choice-Similarity was set to the range of (0.4, 0.6), and $\text{RANGE}_{\text{context}}$ in Context-Similarity to (0.5, 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 compute the similarity between events, we used word vectors that were induced from 200 million sentences of the Japanese web corpus using word2vec.[10] The length range $\text{RANGE}_{\text{length}}$ was set to the range of (0.5, 2.0).

As a result of generation, 103,907 problems were generated from the 104,266

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

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

Table 2.4: Statistics of the constructed dataset.

verified pairs. Table 2.5 provides examples of the generated problems with BERT's predictions described in Section 2.5.1. On this default setting, the mean and median numbers of the eligible candidates of distractors were 3,459 and 1,355, respectively.

Investigation of Human Accuracy To investigate human performance on the generated problems, we randomly sampled them and collected answers by crowdsourcing. Specifically, we prepared 3 sets of 500 problems and performed crowdsourcing on different dates to be answered by different sets of crowdworkers. We collected answers from five crowdworkers per problem. As a result, the average accuracy of individual crowdworkers was 83.8%, and the accuracy of answers aggregated by majority voting was 88.9%.

2.4.5 Building a Japanese Commonsense Contingency Reasoning Dataset

We built a dataset from the generated problems by splitting them into training, development, and test splits with the ratio 8:1:1. In order to reduce leakage, we make these splits so that the pair of core events constituting a base of each problem does not overlap between the training and development/test splits.[11] The statistics of the constructed dataset are organized in Table 2.4.

[11] For instance, the base of the problem in Figure 2.1 is "I'm hungry, so I have a meal" and composed of the pair of core events "be hungry → have a meal". If the training split contains this problem, it means that the development/test split does not contain problems generated from bases that are composed of the same pair of core events like "I'm hungry, so I have a meal at a restaurant".

Comparison to Existing Similar Datasets

We investigate whether our constructed dataset includes some knowledge that is not included so much in existing commonsense contingency reasoning datasets. In this study, we chose SWAG (Zellers et al., 2018) and Social IQA (Sap et al., 2019) for comparison. This is because the two datasets are deemed relevant to our constructed dataset as they focus on contingent relation between phrases or clauses and are similar in size.

As described in Section 2.2.2, SWAG is based on video captions; thus, it primarily focuses on contingent relation between actions and may not include knowledge about states or emotions induced from context[12] so much. To confirm this, we investigated the percentage of examples where the correct choice contains an adjective phrase.[13] As a result, the percentages in SWAG and our constructed dataset were 1.1% and 15.7%, respectively.

Social IQA is a commonsense contingency reasoning dataset consisting of 38k three-choice questions created based on ATOMIC (cf. Section 1.3.2). It covers knowledge about contingent relations between events and related mental states (e.g., intents and emotions). On the other hand, we presume it may not include knowledge about negative condition[14] so much because most of the event expressions in ATOMIC are in basic form. To confirm this, we investigated the percentage of examples where the correct choice is in negative form. As a result, the percentages in Social IQA and our constructed dataset were 1.6% and 11.3%, respectively. These results suggest that our constructed dataset includes a broader range of knowledge in the sense that it contains more diverse adjective clauses and negation expressions.

2.5 Experiments

We conducted experiments to investigate the accuracy of a high-performance language model on the constructed dataset.

[12] For instance, “hang out with friends → be fun”.

[13] We used Stanza (Qi et al., 2020) for constituency parsing.

[14] For instance, “be sleepy → not get motivated”.

2.5.1 Model

We employed the BERT model for experiments. BERT (Devlin et al., 2019) is one of the general-purpose language models, which has achieved high performance on various NLP tasks such as NLI, shallow question answering, and so forth. In order to apply the model to each downstream task, a linear layer is added on top of the output, and all the model parameters are fine-tuned on the task.

For a pre-trained model of BERT, we adopted the pre-trained Japanese BERT$_{\text{LARGE}}$ WWM model,[15] which is pre-trained on 18 million sentences of Japanese Wikipedia with the whole word masking strategy.

2.5.2 Experimental Settings

The task is to choose the most appropriate sentence as the continuation of a given context from four choices, as illustrated in Figure 2.1. The score of each choice is computed by feeding a pair of a context and the choice delimited by special tokens referring to the previous work (Talmor et al., 2019). For instance, the context "お腹 が 空いた ので (I'm hungry, so)" and the choice "ご飯 を 食べる (I have a meal)" become "[CLS] お腹 が 空いた ので [SEP] ご飯 を 食べる [SEP]". The hidden representation of each [CLS] token is converted into a scalar through an added linear layer, which is regarded as the score of each choice.

During the training phase, we define the following objective function:

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

where N is the number of training examples, j is the index of a correct choice among 1 to 4, s_{ki} is the score of the i-th choice of the k-th example.

During the evaluation phase, we regard the choice with the highest score as an answer by a computer. We evaluated the model by accuracy.

The major hyper-parameters are as follows: epoch of 3, maximum sequence length of 128, a batch size of 8,[16] and a learning rate of 2e-5.

[15] `https://nlp.ist.i.kyoto-u.ac.jp/index.php?ku_bert_japanese`

[16] If a batch size is set to N, it means we input N problems at the same time; thus, the input to a model consists of 4N input sequences.

Model	Setting	Accuracy
Chance rate		25.0
BERT$_{\text{LARGE}}$		76.0
Human	1 worker	83.8
	5 workers	88.9

Table 2.5: Experimental results on the constructed dataset.

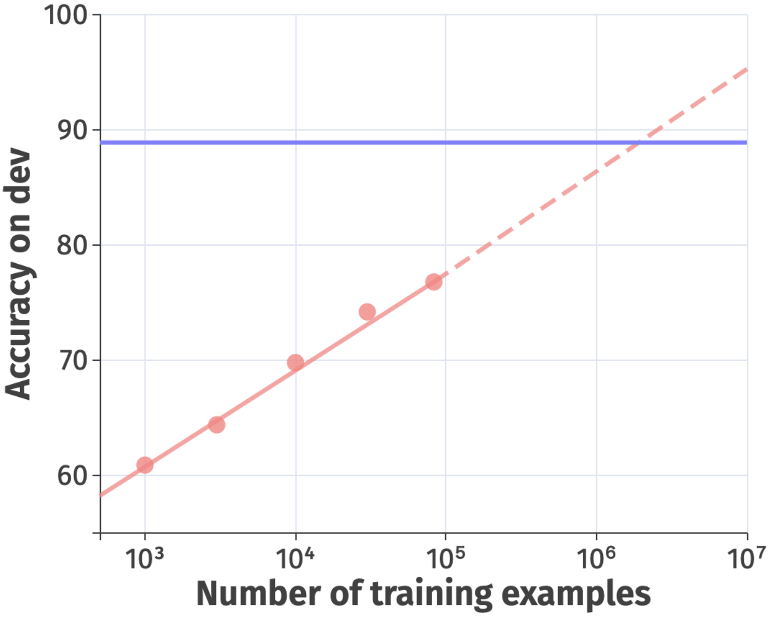


Figure 2.4: Learning curve of the BERT model on the development split.

2.5.3 Experimental Results

BERT$_{\text{LARGE}}$ achieved the accuracy of 76.0 as shown in Table 2.5. It is observable that there is a reasonable performance gap between the NLP model at the time and humans.

Figure 2.4 illustrates the learning curve of the BERT model on the development split. It can be expected by extrapolation that the BERT model requires approximately 1.9 million training examples to achieve human performance, which is not practical. It is meaningful to develop a superior language model to solve

Correct	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 $_{-14.4}$ (treatment is necessary 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 dare not watch TV programs that look interesting) d. 供給も可能かもしれません $_{-14.6}$ (I may be able to provide it)
Incorrect	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 $_{-9.4}$ (dietary considerations, such as mineral and water intake, are necessary)

Figure 2.5: Examples that the BERT model answered correctly and incorrectly. ✓ and × denote the correct choice and the prediction of the BERT model, respectively. The number at the end of each choice represents an output score ($\in [-15, 15]$).

this dataset toward human performance.

2.5.4 Qualitative Analysis

We briefly analyze the predictions by the BERT model. Figure 2.5 provides some examples that the BERT model answered correctly and incorrectly. There were a number of noticeable examples that the BERT model answered incorrectly as a result of overemphasizing lexical overlap between a context and a choice.

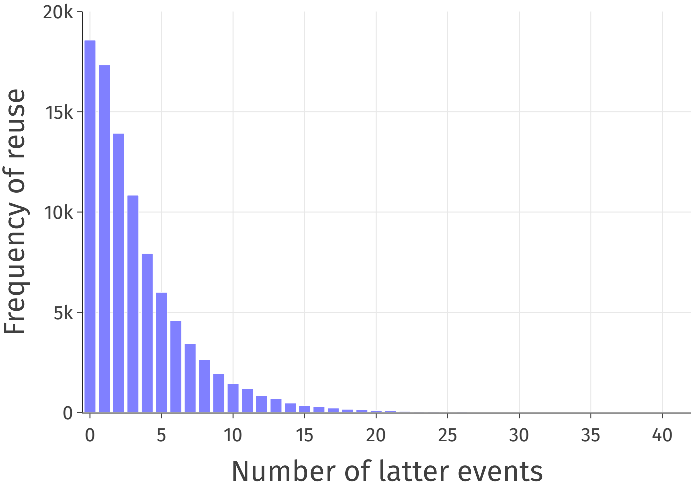


Figure 2.6: Counts of how many times each latter event is used as a distractor.

2.5.5 Investigation of Biases

Several studies have reported that, due to unintended biases in a dataset, some problems can be solved by just observing part of context/question sentences (Gururangan et al., 2018; Zellers et al., 2019; Tamborrino et al., 2020). To investigate the existence of bias in our constructed dataset, we evaluated model performance when feeding only choices during both the training and evaluation phases. For this investigation, we used the same model and hyper-parameters as described in Section 2.5.1.

As a result, $\text{BERT}_{\text{LARGE}}$ achieved an accuracy of 41.2%. Compared with the experimental result in Section 2.5.3, the performance is significantly low, which indicates that our constructed dataset contains low bias.

To investigate the reason why 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 2.6 illustrates the counting result, which indicates some latter events are frequently reused. Thus, we generated problems so as not to use each latter event more than five times as a distractor and evaluated model performance on a dataset built from the generated problems in

$\mathbf{RANGE_{choice}}$	$\mathbf{RANGE_{context}}$	$\mathbf{BERT_{LARGE}}$	**Human**
(0.4, 0.6)	(0.5, 0.7)	76.8	88.9 (83.8)
(0.4, 1.0)	(0.5, 0.7)	72.7	82.2 (78.8)
(0.4, 0.6)	(0.5, 1.0)	73.0	81.8 (77.7)
(-1.0, 0.6)	(0.5, 0.7)	76.7	88.7 (83.9)
(0.4, 0.6)	(-1.0, 0.7)	84.6	92.8 (88.8)

Table 2.6: Investigation results of the conditions on choosing distractors. The numbers in parentheses at the rightmost column represent the average accuracies of individual crowdworkers.

the same manner as described in Section 2.4.5. As a result, $\mathrm{BERT_{LARGE}}$ achieved an accuracy of 30.0%, which suggests that some distractors are easily detected as incorrect due to their high frequency of reuse, leading to higher accuracy than the chance rate.

2.5.6 Investigation of the Conditions on Choosing Distractors

We investigated how the conditions on choosing distractors affect the quality of a dataset. Specifically, we built datasets by removing the upper or lower bounds of each similarity range, $\mathrm{RANGE_{choice}}$ or $\mathrm{RANGE_{context}}$, and evaluated model and human performance on each dataset. We evaluated model performance on each development split using the same model and hyper-parameters as described in Section 2.5.1. We calculated human performance in the same manner as described in Section 2.4.4.

Table 2.6 organizes the investigation results, which indicate the effectiveness of the upper and lower bounds. Specifically, by removing the upper bound, some problems contained distractors that were remarkably similar to the correct choice, and thus, neither the model nor humans could solve them. By removing the lower bound, the relevance between a context and distractors decreased, and thus, the generated problems became easy to solve, especially for the model. Accordingly, it is important to choose moderately similar distractors.

2.5.7 Summary of This Chapter

We proposed a semi-automatic (scalable and low-cost) method for building a commonsense contingency reasoning dataset, which combines automatic extraction from a raw corpus and crowdsourcing. According to the proposed method, we successfully built a Japanese commonsense contingency reasoning dataset comprising 104k multiple-choice questions. As a result of experiments, we demonstrated the reasonable performance gap between the NLP model at the time and humans. We also confirmed that the constructed dataset contained negligible bias, which suggests it can be utilized as a benchmark for further research.

Future work includes improving the quality of the constructed dataset. For instance, as the proposed method automatically generates commonsense contingency reasoning problems from contingent basic event pairs, it does not guarantee that the generated problems can be answered. Furthermore, we could not completely exclude low-quality and noisy sentences that are often found in web text despite filtering by crowdsourcing. In order to solve these issues, it is deemed necessary to manually modify and sort out problems (e.g., by crowdsourcing).

Regarding the acquisition of commonsense knowledge from text, we need to address an essential issue that commonsense knowledge is rarely transcribed due to reporting bias (Gordon and Van Durme, 2013). Toward the acquisition of a broader range of commonsense knowledge, it is worth attempting to apply our proposed method to text intentionally transcribed about our world, such as video captions utilized in SWAG.

Chapter 3

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

3.1 Introduction

In Chapter 2, we successfully built a Japanese commonsense contingency reasoning dataset that can be utilized as a benchmark. The typical next step is to improve model performance on the dataset. In this chapter, we present our work on improving commonsense contingency reasoning by pseudo-data and its application to the related tasks.

Contingency is the discourse relation between events established when one is likely to cause the other. We humans infer contingency on a daily basis. For instance, when reading text, we unconsciously infer what happens next to deepen our understanding. While having a conversation, we guess the next topic from the utterance of the interlocutor to make a contextual and natural response. Thus, 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 3.1: Example from KUCI (English translated version). KUCI is a Japanese QA dataset comprising 104k multiple-choice questions that ask basic contingency directly. ✓ denotes the correct choice.

the linguistic capability to infer contingency is crucial for natural language understanding (NLU).

Recently, language resources regarding contingency have been actively constructed (Roemmele et al., 2011; Mostafazadeh et al., 2016; Zellers et al., 2018; Sap et al., 2019; Sakaguchi et al., 2020). These language resources focus on basic events and evaluate certain kind of commonsense reasoning ability. Although the fundamental linguistic capabilities of computers, such as natural language inference and shallow question answering, have greatly improved with the remarkable development of deep learning, several studies have also empirically demonstrated they still have difficulty in commonsense contingency reasoning (Sap et al., 2019; Sakaguchi et al., 2020; Talmor et al., 2021).

In this study, we set two objectives to validate the importance of contingency reasoning: to improve commonsense contingency reasoning and to investigate the generality of knowledge about basic contingency on the related tasks. To these ends, we utilize the Kyoto University Commonsense Inference dataset (KUCI), the constructed dataset in Chapter 2. KUCI is a Japanese QA dataset comprising 104k multiple-choice questions that ask basic contingency directly. An example is shown in Figure 3.1. This dataset is also characterized by its semi-automatic data construction method: automatic extraction of pairs of basic event expressions that have contingent relation (hereafter, contingent basic event pairs) from a web corpus, verification through crowdsourcing, and automatic generation of commonsense contingency reasoning problems.

It is shown that there is a performance gap between computers and humans on this task. Furthermore, through qualitative analysis, it has been confirmed computers sometimes provide incorrect answers to problems that ask quite basic contingency. A straightforward approach to alleviating the above issue is to manually expand the training data (Hestness et al., 2017; Kaplan et al., 2020; Rosenfeld et al., 2020; Henighan et al., 2020; Bahri et al., 2021). However, it is not practical from a cost perspective to increase the number of training examples manyfold even using crowdsourcing.

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

The second objective of this study is to investigate the generality of knowledge about basic contingency on the related tasks. On the premise that contingency reasoning is crucial for NLU, it can be expected that knowledge about basic contingency 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., 2019; 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 knowledge about basic contingency through quantitative evaluation by performing transfer learning from a commonsense contingency reasoning task to the related tasks.

In summary, we work on improving commonsense contingency 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. Thanks to pseudo-problems, a high-performance language model has achieved near human-level performance on the commonsense contingency reasoning task. We also investigate the transferability of knowledge about basic contingency to the related tasks. Experimental results demonstrate that intermediate-task training on KUCI with pseudo-problems positively affects Japanese Discourse Relation Recognition, the Japanese Winograd Schema Challenge, and the JCommonsenseQA, which suggests the importance of contingency reasoning in NLU.

3.2 Related Work

Thanks to pre-training on large-scale raw corpora, pre-trained language 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 a number of approaches to improving the performance on commonsense reasoning tasks.

Approach to Improving Commonsense Reasoning

One group of approaches is to utilize automatically generated 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 this approach requires the manually constructed language resource, ConceptNet. Staliunaite et al. (2021) proposed a data augmentation method for COPA and its extension (Roemmele et al., 2011; Kavumba et al., 2019), which are summarized as three steps: filtering of web text by several conditions, extraction of causal pairs of clauses with the clue of discourse connectives, and generation of distractors from a language model. They have not investigated its 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 automatically generated examples that imitate WSC.

Transferability of Commonsense Knowledge

Regarding the second objective of this study, there have been several studies on the transferability of commonsense knowledge from existing language resources. For instance, it has been reported that intermediate-task training on the two benchmarks for commonsense contingency reasoning, Social IQA (Sap et al., 2019) 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 language resources for commonsense reasoning to examine their interactions. We investigate the transferability of knowledge about basic contingency in the non-English language, Japanese.

3.3 Approach

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

3.3.1 Method of Generating Pseudo-Problems

The construction method of KUCI consists of the following four steps (cf. Figure 3.2):

1. Acquire high-frequency predicate-argument structures (**core events**) from case frames (Kawahara and Kurohashi, 2006; Kawahara et al., 2014).

2. Extract **contingent basic event pairs**, event pairs that are unambiguously connected by some discourse connectives representing contingent relation and composed of core events, from parsed text.

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

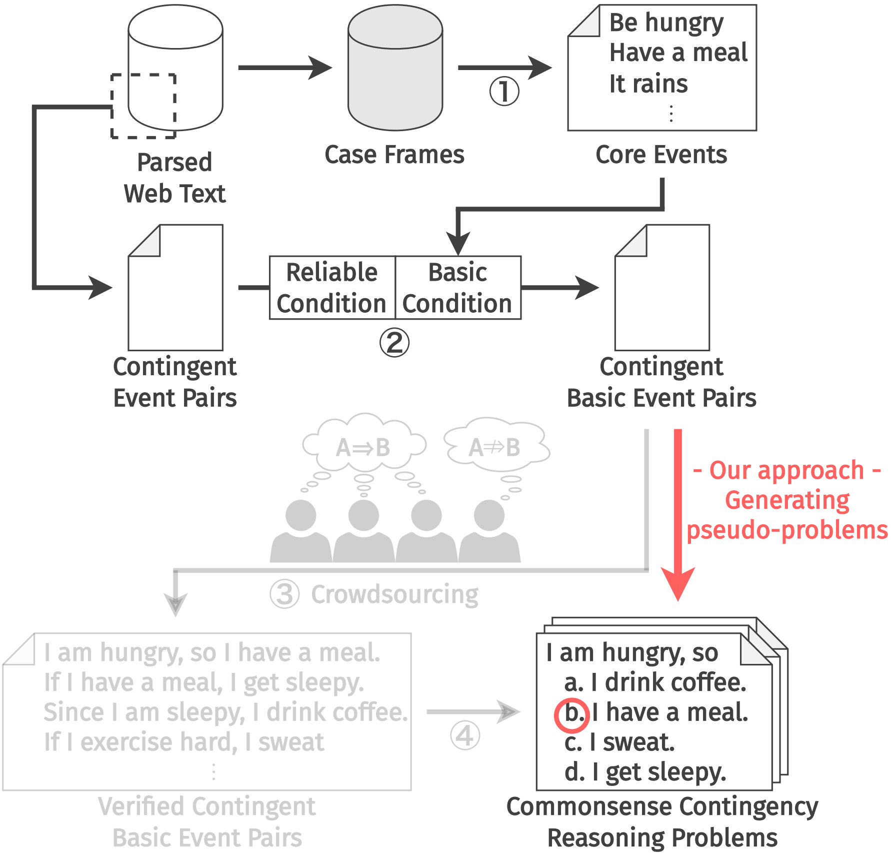


Figure 3.2: Overview of the method of generating commonsense contingency reasoning problems in KUCI (gray) and pseudo-problems (red). Details are described in Section 2.3.

4. Generate problems by taking one of the verified pairs (**base**) and choosing distractors from the latter events of other pairs that are moderately similar to the base.

In the above procedure, it becomes possible to automatically generate pseudo-problems that imitate commonsense contingency reasoning problems by omitting step 3. For the parameters in the method, such as the thresholds of frequency for acquiring core events and the conditions on choosing distractors, we set them to the same values as in the construction of KUCI (cf. Section 2.4).

3.3.2 Extraction of Contingent Basic Event Pairs

According to the method described in Section 3.3.1, we automatically extracted contingent basic event pairs from a Japanese web corpus comprising 3.3 billion sentences. This corpus is also part of an in-house corpus that has been constructed by crawling web text from 2006 to 2015, but there is no overlap of sentences between it and the corpus used for the construction of KUCI. As a result, we extracted 915k contingent basic event pairs. Considering one-third of the extracted event pairs were removed by crowdsourcing as reported in Section 2.4.3, we expect about 600k event pairs to be valid.

3.3.3 Dealing with Data Leakage

There is a potential issue with generating training data from large-scale raw corpora, which is called "Data Contamination" (Brown et al., 2020; Elazar et al., 2023). This issue is that raw corpora 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.[1] 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 a pair of core events that also constitutes any base in evaluation data.

For instance, the base of the problem in Figure 3.1 is "I'm hungry, so → I have a meal at a family restaurant" and composed of the pair of core events "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}, the length of which (10) exceeds 75% of the word

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

count of the base (11). It is also composed of the same pair of core events; thus, it is excluded by both filters.

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

3.3.4 Generation of Pseudo-problems

We proceeded to automatically generate pseudo-problems. As a result, we obtained 862k pseudo-problems from the 881k unverified pairs. The number of pseudo-problems is about ten times as large as that of the training examples in KUCI (83k).

To investigate the quality of pseudo-problems, we randomly sampled 100 problems and manually evaluated them. As a result, 71 out of 100 problems were judged as answerable, which appears to be sufficient quality for automatically generated data.

3.4 Experiments

We conducted experiments to investigate the effectiveness of incorporating pseudo-problems into training in a commonsense contingency reasoning task and the related tasks.

3.4.1 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).[2] It is pre-trained on the full text of Japanese Wikipedia for 1.1 million steps with a batch size of 4,096. It has been reported that the performance is relatively high among the pre-trained Japanese BERT models owing to partly referring to the

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

pre-training configuration of RoBERTa (Liu et al., 2019). The model architecture is the same as BERT$_{\mathrm{BASE}}$.

XLM-R We adopted the XLM-RoBERTa$_{\mathrm{LARGE}}$ model,[3] which is pre-trained on a huge multilingual corpus consisting of Wikipedia and CC-100 (Wenzek et al., 2020). The model architecture is the same as BERT$_{\mathrm{LARGE}}$, but the embedding layer is relatively large due to its multilingual vocabulary. It was one of the high-performance publicly available pre-trained language models for Japanese at the time.

3.4.2 Experimental Settings

The hyper-parameters used in the experiments are included in Appendix A.1.

Commonsense Contingency Reasoning Task

As mentioned in Section 3.1, we utilized KUCI for evaluating the linguistic capability to infer basic contingency. The task is to choose the most appropriate sentence as the continuation of a given context from four choices. KUCI contains 83,127/10,228/10,291 examples in training/development/test split, respectively.

During the training 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 feeding pairs of a context and the choice delimited 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 contingency reasoning problems and pseudo-problems. The above can be expressed by the following equations.

$$H = -\frac{1}{N}\sum_{k=1}^{N} \log \frac{\exp(\mathbf{s}_{kj})}{\sum_{i=1}^{4} \exp(\mathbf{s}_{ki})}$$

$$L = H_{ccr} + \lambda \times H_{pseudo}$$

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

where N is the number of training examples, 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 contingency reasoning problems or pseudo-problems, and λ is the weight for pseudo-problems.

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

Method for Comparison To investigate the effectiveness of a multiple-choice format, we compared it to an additional pre-training method referring to Task-Adaptive Pre-Training (Gururangan et al., 2020). Specifically, we performed an additional Masked Language Modeling (MLM) task on the 881k unverified pairs used for generating pseudo-problems and then fine-tuned the additionally pre-trained model on a target task. For convenience, we name it “AMLM”.

Related Tasks

To investigate the generality of knowledge about basic contingency, we conducted transfer learning from a commonsense contingency reasoning task to the related tasks. In this study, we employed Japanese Discourse Relation Recognition, Japanese Winograd Schema Challenge, and JCommonsenseQA as the related tasks.

Japanese Discourse Relation Recognition Discourse Relation Recognition (DRR) is the task of identifying discourse relations between clauses. In addition to contingent relation, this task requires an understanding of various discourse relations such as “Purpose” and “Concession”.

We used the Kyoto University Web Document Leads Corpus (KWDLC)[4] (Kawahara et al., 2014; Kishimoto et al., 2018, 2020) for this task. KWDLC has been built by collecting the first three sentences of various kinds of crawled web text, the size of which amounts to 6,445 documents. All the documents have been annotated with discourse relations between clauses by crowdsourcing. Furthermore, 500 out of 6,445 documents have also been annotated by linguistic

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

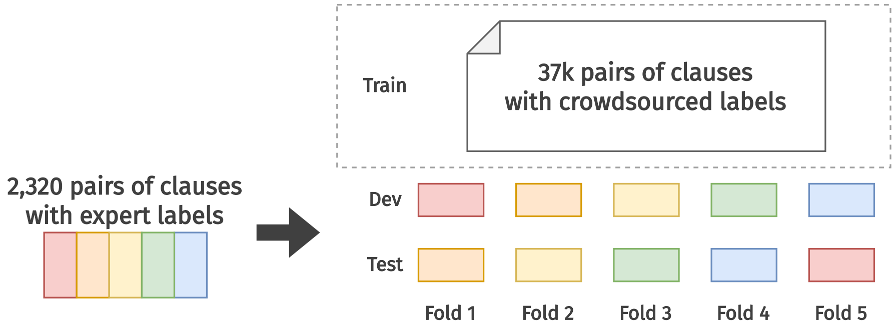

Train
37k pairs of clauses with crowdsourced labels
2,320 pairs of clauses with expert labels
Dev
Test
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5

Figure 3.3: Illustration of five-fold cross-validation on KWDLC. We excluded some pairs of clauses from training data for each fold to ensure that there is no overlap of pairs of clauses between the training and development/test data.

experts. In this study, we used 37k pairs of clauses with crowdsourced labels for training data and evaluated the classification performance on 2,320 pairs of clauses with expert labels using five-fold cross-validation (cf. Figure 3.3).

The task is formulated as a seven-way classification of discourse relations given a pair of clauses, including “No Relation”. We fine-tuned models according to the sentence-pair classification framework proposed by Devlin et al. (2019). We adopted micro-averaged precision, recall, and F1 score computed without examples with the “No Relation” label as evaluation metrics.

Japanese Winograd Schema Challenge Winograd Schema Challenge (WSC) is the task of choosing the antecedent of a pronoun from two candidates (Levesque, 2011). The task itself is coreference resolution but designed to require commonsense reasoning. JWSC[5] (Shibata et al., 2015) has been built by translating the Rahman and Ng (2012) version of WSC into Japanese. Here is an example:

(1) ライオンはシマウマを食べる．それ（ライオン／シマウマ）は捕食動物だからだ．
(Lions eat zebras because they are predators.)

As shown in the above example, JWSC contains a number of questions that ask

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

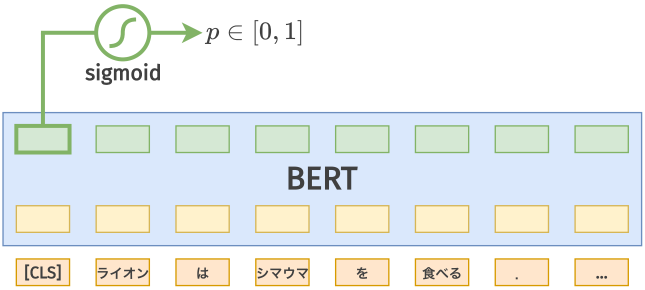


Figure 3.4: Illustration of a BERT-based logistic regression classifier.

basic contingency indirectly.

As we excluded event pairs that contain demonstratives for quality control (cf. Section 2.3.2), there is concern that intermediate-task training on KUCI with pseudo-problems might hurt performance on JWSC due to forgetting knowledge about demonstratives. Accordingly, we recast JWSC as binary question answering by substituting a pronoun with each antecedent candidate. The resulting dataset is balanced and consists of 2,644/1,128 examples for training/test split, respectively. As the development split is not provided, we ran five-fold cross-validation by splitting the training split into 8:2. We trained BERT-based logistic regression classifiers (cf. Figure 3.4) and evaluated them by accuracy and Area Under the ROC Curve (AUC).

JCommonsenseQA JCommonsenseQA (JCQA)[6] (Kurihara et al., 2022) is the Japanese version of CommonsenseQA (Talmor et al., 2019) and consists of 11k five-choice questions regarding a broad range of relations between basic concepts. Each question is manually created based on a subgraph extracted from ConceptNet (Speer et al., 2017) by crowdsourcing. As commonsense contingency reasoning problems and pseudo-problems are composed of basic events, models learn co-occurrence of basic phrases through training on them, which is expected to be beneficial to JCQA. We fine-tuned and evaluated models according to the same method described in Section 3.4.2 as the task is multiple-choice question answering.

[6] `https://github.com/yahoojapan/JGLUE/tree/v1.0.0/datasets/jcommonsenseqa-v1.0`

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

Table 3.1: Experimental results on the commonsense contingency reasoning task. 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.

3.4.3 Experimental Results

Commonsense Contingency Reasoning Table 3.1 shows the experimental results on the commonsense contingency reasoning task.[7] Thanks 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.

[7] We also conducted a preliminary investigation of the performance of the pre-trained Japanese RoBERTa$_{\text{LARGE}}$ model (`https://huggingface.co/nlp-waseda/roberta-large-japanese`), which had been released after the completion of these experiments. As a result, the accuracy on the "KUCI" and "KUCI + Pseudo-problems ($\lambda = 0.5$)" settings was 90.0 ± 0.1 and 90.5 ± 0.5, respectively. The performance on some related tasks was also comparable to human performance without transfer learning; thus, we omit discussion of the effect of transfer learning on RoBERTa$_{\text{LARGE}}$ in this thesis.

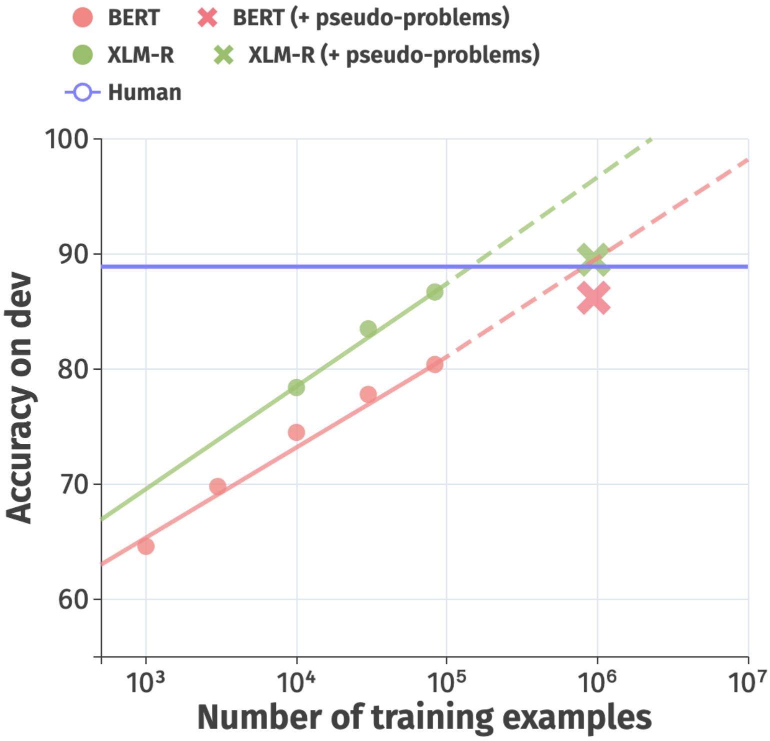


Figure 3.5: Learning curves of the BERT and XLM-R models on the development split of KUCI. We excluded degeneration results of the XLM-R model when fine-tuned on a small number of training examples ($N \in \{10^3, 3 \times 10^3\}$).

Figure 3.5 illustrates the learning curves of the BERT and XLM-R models on the development split of KUCI. The crosses 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.

Japanese Discourse Relation Recognition Regarding JDRR, it is observable from Table 3.2 that intermediate-task training on KUCI with pseudo-problems is effective in discourse relation recognition, especially in BERT. As these problems are based on contingent basic event pairs, which are connected by some discourse connectives representing causal or conditional relation (cf. Section 2.4.2), we presume knowledge about these discourse relations is successfully transferred.

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

Table 3.2: Experimental results on the Japanese discourse relation recognition task. The scores are the mean and standard deviation over three runs of five-fold cross-validation with different random seeds. As with Table 3.1, arrows denote multi-stage fine-tuning. Note that we performed additional Masked Language Modeling (AMLM) on the 881k unverified pairs used for generating pseudo-problems, not the training examples in KWDLC, in order to compare how to utilize pseudo-problems. Human performance is calculated using 500 documents that are annotated by both experts and crowdworkers, with expert annotation as ground truth and crowdsourced annotation as predictions.

Table 3.3 organizes the detailed results on the Japanese discourse relation recognition task. The BERT and XLM-R models transferred from KUCI with pseudo-problems perform better on classifying causal and purpose relations. Compared with crowdworkers, there is room for improvement in the precision of concession and infrequent relations.

Model	Setting	Ca./Re.	Cond.	Purp.	Just.	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 3.3: Detailed results of ensemble models on the Japanese discourse relation recognition task. The third to eighth columns stand for the discourse relations, "Cause/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 the total numbers of true positives and false positives.

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

Table 3.4: Experimental results on JWSC. The scores are the mean and standard deviation over three runs of five-fold cross-validation with different random seeds. † denotes that the result includes a few degenerate runs. We also report the result excluding the degenerate runs in parentheses for reference. Regarding the "AMLM → KUCI → JWSC" setting of XLM-R, the model failed to learn.

Japanese Winograd Schema Challenge The experimental results on JWSC are shown in Table 3.4. We observed a few degenerate runs[8] (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 a similar 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;[9] thus, we consider putting much emphasis on contingent relation may rather worsen performance. Learning various discourse relations is a promising solution, which we leave for future work.

[8] The training runs that a model results in around chance performance. Specifically, we regard less than 0.55 accuracy or AUC as the degenerate runs.

[9] e.g., "James asked Robert a favor. However, James/Robert declined."

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

Table 3.5: Experimental results 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 (Kurihara et al., 2022) (the numbers in the parentheses) for reference.

JCommonsenseQA Referring to Table 3.5, we can see the solid performance gain regarding XLM-R. We presume 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 appear to be somewhat insufficient for adapting the model pre-trained only on Wikipedia (i.e., BERT) to the web domain but effective as auxiliary data for the model pre-trained on CC-100 (i.e., XLM-R).

Comparison to AMLM Although AMLM is somewhat effective in KUCI, it is poor at transferring the knowledge to the related tasks.[10] It can be inferred that the BERT and XLM-R models learn task-specific knowledge.

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

霧が晴れると、	午後から病院へいくので
(When a fog clears,)	(I'm going to see a doctor this afternoon, so)
✓ a. 景色が素晴らしい	a. 滅多に病院に行かない
(the scenery is amazing)	(I rarely see a doctor)
× b. 川の音がすごい	b. 土日は勉強に勤しみます
(the sound of river is loud)	(I'll study hard on weekends)
c. 雪遊びも楽しそうだ	✓ c. 今日は休暇をとる
(playing in the snow sounds nice)	(I take a vacation today)
d. 写真写りがいまいちだ	× d. 火曜日は眠い
(it's not photogenic)	(I'm sleepy on Tuesday)

Figure 3.6: Examples of problems that the BERT model became able to answer correctly by incorporating pseudo-problems into training. ✓ and × denote the correct choice and the choice that BERT previously selected, respectively.

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

Table 3.6: Confusion matrix organizing the numbers of correct and incorrect answers on the development split of KUCI. This matrix shows the results of the BERT model (ensemble).

3.4.4 Qualitative Analysis

Figure 3.6 illustrates examples of problems that BERT became able to answer correctly by incorporating pseudo-problems into training. We can see the improvement in the accuracy of these problems regarding quite basic contingency. The model sometimes gave low scores to all the choices and appeared to select a choice by elimination, which we observed became less frequent. We speculate that pseudo-problems complement the lack of coverage of the training examples in KUCI. For further information, we include the confusion matrix in Table 3.6.

The improvement is greater though the model got to make a wrong prediction to some problems.

3.5 Summary of This Chapter

We improved commonsense contingency reasoning by incorporating large-scale pseudo-problems into training. We automatically generated 862k pseudo-problems from a Japanese web corpus comprising 3.3 billion sentences utilizing the scalability of the construction method of KUCI. Thanks to pseudo-problems, a high-performance pre-trained language model has achieved near human-level performance on the commonsense contingency reasoning task.

We also investigated the effectiveness of learning knowledge about basic contingency in the related tasks: Japanese Discourse Relation Recognition, Japanese Winograd Schema Challenge, and JCommonsenseQA. Experimental results demonstrated that intermediate-task training on KUCI with pseudo-problems has a positive impact on the related tasks, which suggests the importance of contingency reasoning in NLU.

Chapter 4

Synthetic Data Generation for Discourse Relation Recognition

4.1 Introduction

In Chapter 3, a high-performance pre-trained language model has achieved near human-level performance on our constructed dataset thanks to pseudo-problems. Thus, we expand our focus from contingency to discourse relations. Specifically, we work on improving Discourse Relation Recognition, the task of identifying the discourse relation given a pair of text spans.

In order to comprehend the meaning of text, it is essential to understand not only the meanings of individual sentences but also the semantic relations between them. Such semantic relations are called **discourse relations**. Automatic recognition of discourse relations has been actively studied due to its applicability to natural language understanding (NLU) (Bhargava and Ng, 2022) and various natural language processing (NLP) tasks (Saito et al., 2019; Tang et al., 2021).

Penn Discourse Treebank (PDTB) (Prasad et al., 2019) is one of the representative corpora regarding discourse relations. This corpus has been built by annotating 2,162 Wall Street Journal articles with discourse relations between adjacent text spans named *arguments*. An example is shown in Figure 4.1; (hereafter, we express an argument pair as *Arg1* and *Arg2*.) The arguments of the example do not contain any discourse connectives, words or phrases that indicate

Arg1: Maggie Thatcher must be doing something right;
Arg2: her political enemies are screaming louder than ever.
Relation: Contingency.Cause+Belief.Reason+Belief
Connective: “because”

Figure 4.1: Example from PDTB. PDTB defines at most three levels of hierarchical discourse relations. In the example, *Relation* is delimited by periods, and top-, second-, and third-level relations are “Contingency”, “Cause+Belief”, and “Reason+Belief”, respectively. Note that the higher the level, the coarser the granularity. In addition, some discourse connectives are assigned to lexicalize the relations. Regarding implicit discourse relations, the annotated connectives are not present in arguments.

some discourse relations, such as “because”. Such examples are called **implicit discourse relations**.

Discourse Relation Recognition (DRR), especially Implicit Discourse Relation Recognition (IDRR), is a long-standing and challenging problem. Even large language models (LLMs), which have achieved unprecedented performance on a variety of NLP tasks, still cannot solve this task in a straightforward manner.[1] In addition to the complexity of DRR itself, the paucity of training data for some error-prone discourse relations makes the problem even more challenging.

A straightforward solution to the aforementioned problem is to increase the number of annotated examples. However, it is not practical due to requiring cautious annotations by experts. Turning our attention to automatic generation of training data, synthetic data generation using language models has achieved some success recently (Puri et al., 2020; Yang et al., 2020; Schick and Schütze, 2021; Liu et al., 2022). There is room for exploration of their generative capabilities to generate argument pairs that have a given discourse relation, although the low few-shot performance of LLMs in DRR is problematic.

[1] We investigated the few-shot performance of GPT-3.5 and GPT-4 in IDRR and confirmed that it is far behind the fine-tuning performance of much smaller language models, which is described in Section 4.3.2.

In this study, we explore synthetic data generation for DRR using an LLM. We first conduct preliminary experiments to confirm the paucity of training data for some error-prone discourse relations. Based on the preliminary results, we propose a method of generating synthetic data for these error-prone discourse relations using an LLM. Specifically, it is summarized as two folds: extraction of confusing discourse relation pairs based on false negative rate and generation of synthetic data focused on resolving the confusion. We demonstrate the performance gain by incorporating the synthetic data into training.

The proposed method has two key points. First, we utilize a confusion matrix for synthesizing effective data. We address the data scarcity problem of some error-prone discourse relations by generating synthetic data based on a confusion matrix.

Second, we devise a method of generating synthetic data. It is probably ineffective to straightforwardly generate synthetic data for DRR using an LLM due to the low few-shot performance. We presume that it is attributed to the number of discourse relations. In other words, it is challenging for an LLM to learn and distinguish numerous discourse relations from few-shot examples. On the other hand, it is relatively easy to learn a single discourse relation from few-shot examples. Thus, we decompose the process of generating synthetic data into two stages so that only a single discourse relation needs to be learned in each stage. Further details are described in Section 4.4.1.

The contributions of this study are summarized as follows:

- We propose an error-driven method of generating synthetic data for DRR using an LLM.

- According to the proposed method, we built synthetic data several times larger than training examples for some error-prone discourse relations.

- We demonstrated the effectiveness of synthetic data in both English and Japanese DRR.

4.2 Related Work

4.2.1 Improving IDRR

As shown in Figure 4.1, PDTB has two major characteristics: discourse relations are defined hierarchically and lexicalized by discourse connectives. A number of previous studies on improving IDRR have exploited these characteristics.

Utilizing Relation Hierarchy This kind of approach has been on the rise recently. For instance, Long and Webber (2022) introduced contrastive learning and utilized the relation hierarchy to choose hard negatives, assuming it is difficult to classify discourse relations that have the same higher-level ones. However, we demonstrate an encoder-only language model such as RoBERTa (Liu et al., 2019) is apt to confuse infrequent discourse relations with frequent ones rather than misclassify discourse relations that have the same higher-level ones (cf. Section 4.3.3). Jiang et al. (2023) also developed the contrastive framework to learn the relation hierarchy and similarity between examples simultaneously, but the same can be pointed out. Wu et al. (2022) showed the effectiveness of learning to generate labels along the relation hierarchy. This method may suffer error propagation from mispredicted top-level discourse relations.

Utilizing Discourse Connectives Several studies have been devoted to learning implicit discourse relations through discourse connectives for some time. For instance, Nie et al. (2019) and Kishimoto et al. (2020) have reported a performance gain by performing an additional pre-training task to predict masked discourse connectives. Other studies such as Xiang et al. (2022) and Zhou et al. (2022) introduced prompt-based learning and utilized annotated discourse connectives as verbalizers. As implicit discourse relations are mentioned without discourse connectives, it is also worth considering methods not relying on discourse connectives.

Other Approaches Xu et al. (2018) introduced active learning to obtain argument pairs that contain omittable discourse connectives (Rutherford and Xue,

2015) for data augmentation. Jiang et al. (2021) performed joint learning of classification and generation, aiming to deepen the model's understanding of discourse relations through generating arguments. To the best of our knowledge, no studies have been conducted on synthetic data generation for IDRR using an LLM.

4.2.2 Synthetic Data Generation for NLP tasks

After the advent of pre-trained language models, an increasing number of studies have attempted to utilize them for synthetic data generation. For instance, Schick and Schütze (2021) synthesized 121k sentence pairs for semantic textual similarity task using GPT-2 XL (Radford et al., 2019) and achieved superior performance with the synthetic data only. Liu et al. (2022) incorporated human-in-the-loop into synthetic data generation for natural language inference task and built a dataset comprising 108k examples using GPT-3 (Brown et al., 2020). In addition, synthetic data generation has been attempted for other NLP tasks, including question answering (Puri et al., 2020), commonsense reasoning (Yang et al., 2020), and so forth. While recent studies lean toward improving few-shot performance with synthetic data (Meng et al., 2023; Dai et al., 2023), we aim to improve fine-tuning performance of encoder-only language models in IDRR considering the relatively low few-shot performance of LLMs.

4.3 Preliminaries

Our proposed method is motivated by preliminary experimental results of English IDRR. We first describe the task settings and preliminary experimental results.

4.3.1 Task Settings

As there are several variations of preprocessing and evaluation protocols regarding PDTB (Kim et al., 2020), we explicate task settings used in our experiments (Section 4.3.2, 4.3.3, and 4.5).

Version of PDTB PDTB has been updated several times over the years. While the previous version (PDTB-2) (Prasad et al., 2008) has been conventionally used so far, the latest version (PDTB-3) (Prasad et al., 2019) has improved in both quantity and quality of annotations. We adopt PDTB-3 taking into account that more annotated examples are available for generating synthetic data.

Label Set Label sets vary by the version of PDTB and the level of discourse relations to classify. We address the fine-grained classification of second-level (L2) discourse relations and follow Kim et al. (2020) to define a label set for the task. Specifically, we formulate IDRR as a 14-way classification using only the labels with more than 100 examples.

Data Partitioning PDTB consists of 25 sections, and we need to partition them to build a dataset. For a fair comparison with previous studies, we adopt the conventional partition introduced by Ji and Eisenstein (2015), where we use sections 2-20, 0-1, and 21-22 as training, development, and test splits, respectively. For convenience, we call it *PDTB dataset*. The statistics of the PDTB dataset are organized in Table 4.1.

Handling of Multi-labeled Examples Regarding multi-labeled examples, we follow a common practice (Ji and Eisenstein, 2015; Qin et al., 2017). Specifically, during the training phase, we convert them into separate examples. During the evaluation phase, a prediction is regarded as correct if it matches one of the labels.[2]

4.3.2 Few-shot Performance of LLMs

Few studies attempted to employ LLMs for IDRR except Chan et al. (2023), which investigated the zero-shot performance of GPT-3.5 on PDTB-2. We also investigated the few-shot performance of GPT-3.5 and GPT-4 (OpenAI, 2023) on PDTB-3.

[2] We found there are two implementations of this. Let us consider the case where a model predicts “A” to an example with the labels “A” and “B”. One implementation overwrites the prediction with “A” and “B”, while the other ignores the label “B” of the example. This may cause discrepancies in the total number of labels among studies. In this study, we confirmed the implementation in a compared method and adopted the former implementation.

Relation	Train	Synthetic Data		Dev	Test
		Unfiltered	LLM-Filtered		
Temporal.Synchronous	435	2,501	1,286	33	43
Temporal.Asynchronous	1,007	-	-	105	108
Contingency.Cause	4,475	-	-	449	406
Contingency.Cause+Belief	159	940	331	13	15
Contingency.Purpose	1,092	-	-	96	89
Contingency.Condition	150	-	-	18	15
Comparison.Concession	1,164	-	-	105	97
Comparison.Contrast	741	-	-	91	63
Expansion.Conjunction	3,586	-	-	299	237
Expansion.Equivalence	254	1,167	771	25	30
Expansion.Instantiation	1,166	-	-	118	128
Expansion.Level-of-detail	2,601	-	-	274	214
Expansion.Manner	615	-	-	28	53
Expansion.Substitution	343	-	-	32	32

Table 4.1: Statistics of the PDTB dataset and synthetic data. Regarding multi-labeled examples, we counted the labels separately. As synthetic data may vary by a model, we show the statistics of the synthetic data generated from the confusion matrix in Figure 4.3 as a representative.

Experimental Settings

As mentioned in Section 4.3.1, we address the 14-way classification of second-level discourse relations. Figure 4.2 shows the prompt template for few-shot learning on the task. We instructed LLMs to generate one of the labels given the definitions of discourse relations (cf. Table 4.2) and demonstrations.

For the LLMs, we employed the snapshots of GPT-3.5 and GPT-4 from June 13th, 2023 (a.k.a “gpt-3.5-turbo-16k-0613” and “gpt-4-0613”). We retrieved K nearest neighbors of a test example from training examples for each discourse relation and used the $K \times 14$ examples as demonstrations referring to Liu et al.

Relation	Definition
Temporal.Synchronous	there is some degree of temporal overlap between the events described by the arguments
Temporal.Asynchronous	one event is described as preceding the other
Contingency.Cause	the situations described in the arguments are causally influenced but are not in a conditional relation
Contingency.Cause+Belief	evidence is provided to cause the hearer to believe a claim
Contingency.Purpose	one argument presents an action that an agent undertakes with the purpose of the goal conveyed by the other argument being achieved
Contingency.Condition	one argument presents a situation as unrealized (the antecedent), which (when realized) would lead to the situation described by the other argument
Comparison.Concession	an expected causal relation is cancelled or denied by the situation described in one of the arguments
Comparison.Contrast	at least two differences between the arguments are highlighted
Expansion.Conjunction	both arguments, which don ' t directly relate to each other, bear the same relation to some other situation evoked in the discourse
Expansion.Equivalence	both arguments are taken to describe the same situation, but from different perspectives
Expansion.Instantiation	one argument describes a situation as holding in a set of circumstances, while the other argument describes one or more of those circumstances
Expansion.Level-of-detail	both arguments describe the same situation, but in less or more detail
Expansion.Manner	the situation described by one argument presents the manner in which the situation described by other argument has happened or been done
Expansion.Substitution	arguments are presented as exclusive alternatives, with one being ruled out

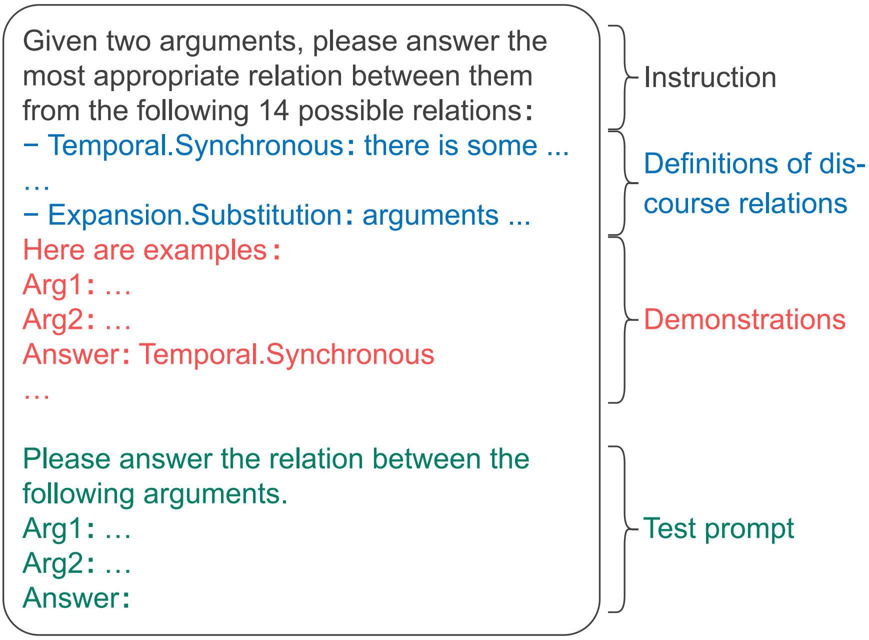
Table 4.2: Definitions of discourse relations in PDTB. They are basically taken from PDTB-3 annotation manual (Webber et al., 2019), but we slightly modify that of "Expansion.Conjunction".

(2022). We made use of the RoBERTa$_{\mathrm{LARGE}}$-based supervised SimCSE[3] (Gao et al., 2021) for retrieving nearest neighbors and set K to 8 considering the token limit of the LLMs. We used the test split of the PDTB dataset for evaluation and evaluated the model by micro-F1 and macro-F1.

Experimental Results

Table 4.3 shows the few-shot performance of GPT-3.5 and GPT-4 on PDTB-3. Despite providing more than 100 examples as demonstrations, the few-shot performance is far behind the fine-tuning performance of the RoBERTa$_{\mathrm{BASE}}$ model.

[3] `https://huggingface.co/princeton-nlp/sup-simcse-roberta-large`



Given two arguments, please answer the most appropriate relation between them from the following 14 possible relations:
- Temporal.Synchronous: there is some ...
...
- Expansion.Substitution: arguments ...

Here are examples:
Arg1: ...
Arg2: ...
Answer: Temporal.Synchronous
...

Please answer the relation between the following arguments.
Arg1: ...
Arg2: ...
Answer:

Figure 4.2: Prompt template for few-shot learning on PDTB-3.

Model	Setting	Micro-F1	Macro-F1
GPT-3.5	few-shot	23.2	19.0
GPT-4	few-shot	29.4	30.9
RoBERTa$_{\text{BASE}}$	Vanilla	64.2	57.1

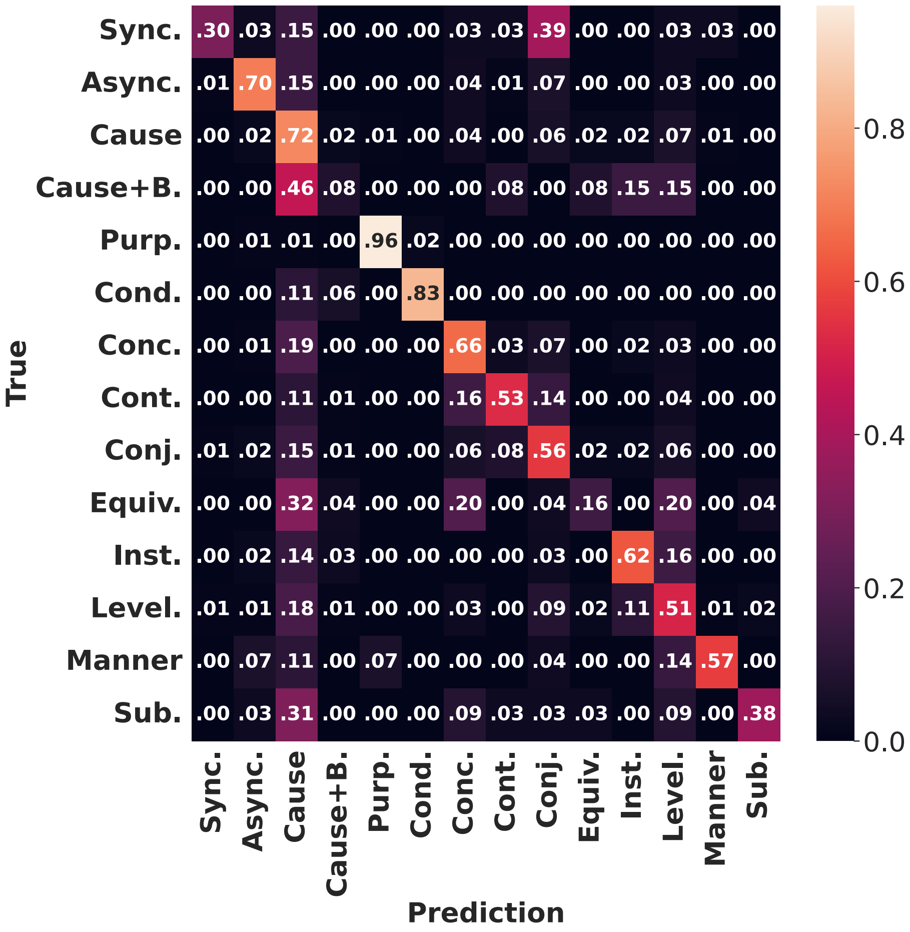
Table 4.3: Experimental results of few-shot learning on PDTB-3. The vanilla fine-tuning performance of RoBERTa$_{\text{BASE}}$ is taken from Table 4.5.

4.3.3 Confusion Matrix of Encoder Model

In order to identify the propensity for error in a commonly used model, we analyzed the confusion matrix.

Experimental Settings

We investigated the confusion matrix of the RoBERTa$_{\text{BASE}}$ model, which has been employed in numerous recent studies. We fine-tuned the RoBERTa$_{\text{BASE}}$



	Sync.	Async.	Cause	Cause+B.	Purp.	Cond.	Conc.	Cont.	Conj.	Equiv.	Inst.	Level.	Manner	Sub.
Sync.	.30	.03	.15	.00	.00	.00	.03	.03	.39	.00	.00	.03	.03	.00
Async.	.01	.70	.15	.00	.00	.00	.04	.01	.07	.00	.00	.03	.00	.00
Cause	.00	.02	.72	.02	.01	.00	.04	.00	.06	.02	.02	.07	.01	.00
Cause+B.	.00	.00	.46	.08	.00	.00	.00	.08	.00	.08	.15	.15	.00	.00
Purp.	.00	.01	.01	.00	.96	.02	.00	.00	.00	.00	.00	.00	.00	.00
Cond.	.00	.00	.11	.06	.00	.83	.00	.00	.00	.00	.00	.00	.00	.00
Conc.	.00	.01	.19	.00	.00	.00	.66	.03	.07	.00	.02	.03	.00	.00
Cont.	.00	.00	.11	.01	.00	.00	.16	.53	.14	.00	.00	.04	.00	.00
Conj.	.01	.02	.15	.01	.00	.00	.06	.08	.56	.02	.02	.06	.00	.00
Equiv.	.00	.00	.32	.04	.00	.00	.20	.00	.04	.16	.00	.20	.00	.04
Inst.	.00	.02	.14	.03	.00	.00	.00	.00	.03	.00	.62	.16	.00	.00
Level.	.01	.01	.18	.01	.00	.00	.03	.00	.09	.02	.11	.51	.01	.02
Manner	.00	.07	.11	.00	.07	.00	.00	.00	.04	.00	.00	.14	.57	.00
Sub.	.00	.03	.31	.00	.00	.00	.09	.03	.03	.03	.00	.09	.00	.38

Figure 4.3: Normalized confusion matrix of the RoBERTa$_{\text{BASE}}$ model. We applied row normalization to the confusion matrix so that each element represents sensitivity or false negative rate.

pre-trained model[4] on the PDTB dataset and calculated a confusion matrix on the development split. Training details and hyper-parameters are described later in Section 4.5.1.

Ground Truth	**Prediction**
Contingency.Cause+Belief	Contingency.Cause
Temporal.Synchronous	Expansion.Conjunction
Expansion.Equivalence	Contingency.Cause
Expansion.Substitution	Contingency.Cause
Expansion.Equivalence	Comparison.Concession

Table 4.4: Top-5 confusing discourse relation pairs in the RoBERTa$_{\text{BASE}}$ model.

Experimental Results

We define the degree of confusion by false negative rate considering the class imbalance as seen in Table 4.1. Figure 4.3 illustrates the normalized confusion matrix of the RoBERTa$_{\text{BASE}}$ model. Several non-diagonal elements indicate a high degree of confusion, i.e., much room for improvement. Furthermore, it is observable from Table 4.4 that RoBERTa$_{\text{BASE}}$ is apt to confuse infrequent discourse relations such as “Cause+Belief” and “Equivalence” with frequent ones rather than misclassify discourse relations that have the same higher-level ones.

4.4 Synthetic Data Generation

Based on the preliminary experimental results, we propose an error-driven method of generating synthetic data for improving fine-tuning performance of an encoder-only language model in DRR.

4.4.1 Proposed Method

The proposed method of generating synthetic data consists of the following three steps (cf. Figure 4.4):

1. Extract top-k confusing discourse relation pairs based on false negative rate.

2. For each confusing discourse relation pair (R_{true}, R_{pred}), retrieve training examples that have R_{true} as the source of synthetic data.

[4] `https://huggingface.co/roberta-base`

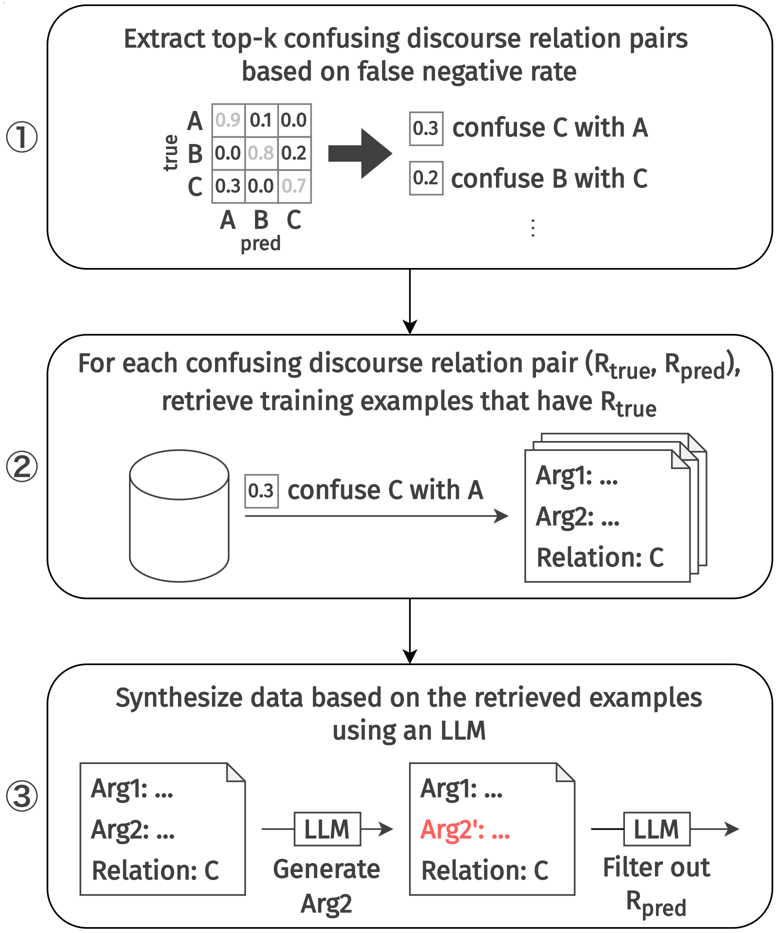

①
Extract top-k confusing discourse relation pairs based on false negative rate
true
A 0.9 0.1 0.0
B 0.0 0.8 0.2
C 0.3 0.0 0.7
A B C
pred
0.3 confuse C with A
0.2 confuse B with C
⋮
②
For each confusing discourse relation pair (R_true, R_pred), retrieve training examples that have R_true
0.3 confuse C with A
Arg1: ...
Arg2: ...
Relation: C
③
Synthesize data based on the retrieved examples using an LLM
Arg1: ...
Arg2: ...
Relation: C
LLM
Generate Arg2
Arg1: ...
Arg2': ...
Relation: C
LLM
Filter out R_pred

Figure 4.4: Overview of the proposed method.

3. Synthesize data based on the retrieved examples using an LLM.

The following paragraphs explicate each step.

Extraction of Confusing Discourse Relation Pairs The first step is to extract confusing discourse relation pairs referring to a confusion matrix. As described in Section 4.3.3, we fine-tune a model, calculate a confusion matrix on the development split, and extract top-k confusing discourse relation pairs based on false negative rate. We utilize false negative rate as the degree of confusion to treat infrequent and frequent discourse relations equally.

First stage

Given two arguments, the relation R_{true} is defined as
⟨ the definition of R_{true} ⟩
Here are examples that have the relation R_{true}:
⟨ demonstrations ⟩
Please write down arguments that have the relation R_{true} to the argument
⟨ Arg1 ⟩.

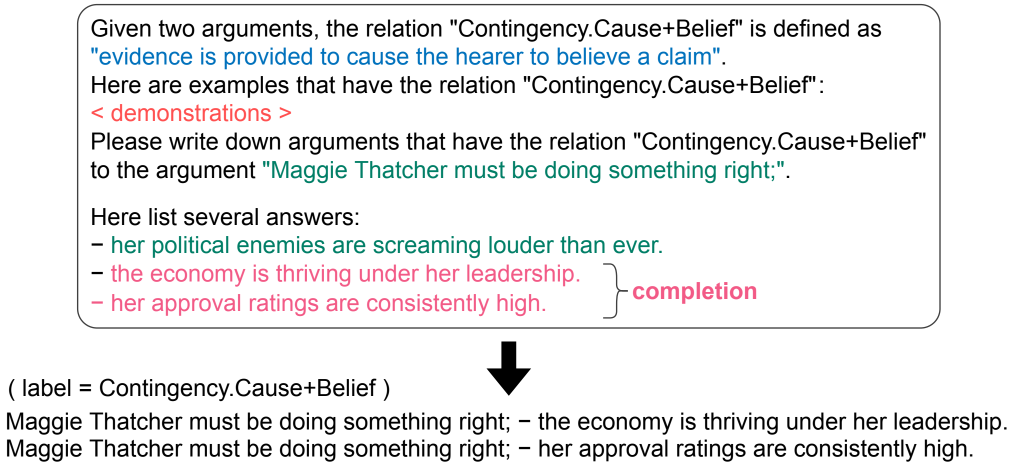
Here list several answers:
- ⟨ Arg2 ⟩
-

Second stage

Given two arguments, the relation R_{pred} is defined as
⟨ the definition of R_{pred} ⟩
Here are examples that have the relation R_{pred}:
⟨ demonstrations ⟩
Please answer whether the two arguments
⟨ pair of Arg1 and synthetic Arg2 ⟩ have the relation R_{pred} or not. An answer must end with “Yes.” or “No.”.

Figure 4.5: Prompt templates for an LLM. We adopt two-stage prompting to generate synthetic data. R_{true} and R_{pred} represent ground-truth and mispredicted discourse relations, respectively.

Retrieval of Training Examples Next, we prepare the source of synthetic data. We utilize training examples judging it is difficult to generate argument pairs that have some discourse relation from scratch. Specifically, for each confusing discourse relation pair (R_{true}, R_{pred}), we retrieve all the training examples that have R_{true} in preparation for the following synthesis process.



Given two arguments, the relation "Contingency.Cause+Belief" is defined as "evidence is provided to cause the hearer to believe a claim".
Here are examples that have the relation "Contingency.Cause+Belief":
< demonstrations >
Please write down arguments that have the relation "Contingency.Cause+Belief" to the argument "Maggie Thatcher must be doing something right;".

Here list several answers:
− her political enemies are screaming louder than ever.
− the economy is thriving under her leadership.
− her approval ratings are consistently high.

(label = Contingency.Cause+Belief)
Maggie Thatcher must be doing something right; − the economy is thriving under her leadership.
Maggie Thatcher must be doing something right; − her approval ratings are consistently high.

Figure 4.6: Illustration of the first stage of synthetic data generation using the example in Figure 4.1. The definitions of discourse relations are taken from PDTB-3 annotation manual[5] (Webber et al., 2019).

Synthesis of Data Finally, we synthesize data focused on resolving the confusion. As mentioned in Section 4.1, we adopt two-stage prompting to synthesize data (cf. Figure 4.5). Specifically, in the first stage, we instruct an LLM to generate a candidate list of Arg2 given Arg1, original Arg2, and the definition of R_{true}. We synthesize Arg2 considering the unidirectionality of language models. Figure 4.6 demonstrates the aforementioned process using the example in Figure 4.1. Synthetic data can be obtained by splitting completion by the item mark "- " and combining each split with Arg1 and the label of R_{true}. In the second stage, we ask an LLM whether each pair of Arg1 and synthetic Arg2 has R_{pred} or not. Regarding the demonstrations for learning R_{true}/R_{pred}, we use K nearest neighbors of a source example referring to Liu et al. (2022), which are retrieved from training examples that have R_{true}/R_{pred}.

4.4.2 Generation of Synthetic Data

According to the proposed method, we generated synthetic data from top-1, 3, and 5 confusing discourse relation pairs to examine the effect of k in later experiments.

[5] `https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf`

We fixed the value of K, the number of nearest neighbors for learning the relation R_{true}/R_{pred}, to 8 referring to Min et al. (2022) to avoid excessive parameter tuning. For an LLM, we employed GPT-4 (a.k.a "gpt-4-0613"). Table 4.1 includes the statistics of the synthetic data generated from top-3 confusing discourse relation pairs for RoBERTa$_{\text{BASE}}$ as a representative.

Analysis of Synthetic Data In order to analyze the quality of synthetic data quantitatively, we sampled 30 examples each for the "Cause+Belief", "Synchronous", and "Equivalence" relations and manually verified them. We chose these three relations because they were top confusing discourse relations in all the experimental settings we tested. As a result of manual verification, 20, 20, and 23 examples of "Cause+Belief", "Synchronous", and "Equivalence" were judged as valid, which appears to be acceptable quality as synthetic data.

We also analyzed the synthetic data qualitatively. "Cause+Belief" is required that one argument expresses some belief, and the other provides its justification. As is the example in Table 4.7, synthetic Arg2 is sometimes factual and inconsistent with original Arg2 when it expresses some belief. One of the possible remedies is to utilize third-level discourse relation to choose examples whose Arg1 expresses some belief.

Regarding "Synchronous", we observed GPT-4 was apt to include discourse connectives such as "while" to establish the relation. Although such examples are valid, this may cause shortcut learning (Geirhos et al., 2020), which raises the need for refining instructions.

Synthetic data of "Equivalence" was often judged as valid. One of the possible reasons is that the discourse relation is regarded as a kind of paraphrasing and is relatively easy to understand for the LLM.

4.5 Experiments on English IDRR

We conducted experiments to examine the effectiveness of incorporating synthetic data into training.

Arg1: A half-hour later, the woman is smiling and chatting;
Original Arg2: the demon seems to have gone.
Synthetic Arg2: her mood has significantly improved.
Relation: Contingency.Cause+Belief

Arg1: ensure the same flow of resources
Original Arg2: and reduce the current deficit.
Synthetic Arg2: while maintaining the current workforce.
Relation: Temporal.Synchronous

Arg1: It's a nervous market.
Original Arg2: It was all over the place.
Synthetic Arg2: The market is highly unpredictable.
Relation: Expansion.Equivalence

Figure 4.7: Examples of synthetic data.

4.5.1 Experimental Settings

Data and Model

We used the PDTB dataset and the synthetic data generated by the proposed method as described in Section 4.3.1, 4.4.1. These statistics are organized in Table 4.8.

We evaluated the performance of the RoBERTa (Liu et al., 2019) model to compare with previous studies. We employed the base-[4] and large-size[6] pre-trained models hosted on Hugging Face Hub.

Training Details

During the training phase, we minimize the standard softmax cross-entropy loss. When incorporating synthetic data into training, we minimize the weighted sum

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

of the losses of training examples and synthetic data, which is expressed by the following equations:

$$H = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_y(x)}}{\sum_{y' \in [Y]} e^{f_{y'}(x)}}$$

$$L = H_{training} + \lambda \times H_{synthetic}$$

where N is a batch size, Y is a set of classes, $f_y(x)$ is the logit for the class y, and λ is the weight for synthetic data.

During the evaluation phase, we evaluate the model by Micro-F1 and Macro-F1. We measure the performance on the development split per epoch and adopt the model parameters with the best dev Macro-F1 for evaluation on the test split.

Compared Methods

We adopted the following methods for comparison.

Vanilla On this setting, we merely fine-tune models without synthetic data.

Logit Adjustment (Menon et al., 2021) Based on the results of Table 4.4, we speculate the synthetic data generated by our proposed method is effective in learning long-tail discourse relations. Thus, we compare logit adjustment with our proposed method as a baseline of learning long-tail classes. This method adjusts logits when computing the standard softmax cross-entropy loss so that the rarer the class, the greater the loss. The above is expressed by the following equation.

$$L = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_y(x)+\tau \log \pi_y}}{\sum_{y' \in [Y]} e^{f_{y'}(x)+\tau \log \pi_{y'}}}$$

where π_y is an estimate of the class prior and τ is the temperature. We used the class frequencies on the training examples as π_y and set τ to 1.0 referring to the authors' report.

Long and Webber (2022) This is one of the state-of-the-art (SOTA) methods for IDRR on PDTB-3. As described in Section 4.2.1, they achieved superior performance by introducing contrastive learning. They also used additional training examples generated by inserting annotated discourse connectives between arguments.

Vanilla-filtered On this setting, we use the vanilla fine-tuned model instead of an LLM for filtering synthetic data.

Hyper-Parameters

Regarding baselines, we performed a grid search of learning rate from {5e-6, 1e-5, 2e-5} and chose the one that achieved the best Macro-F1 on the development split. When incorporating synthetic data into training, we used the same hyper-parameters but performed a grid search of λ, the weight for synthetic data, from {0.5, 0.25}. As we generated synthetic data from top-1, 3, and 5 confusing discourse relation pairs, we adopted the one that achieved the best Macro-F1 on the development split. Specifically, we used the synthetic data generated from top-3 and top-5 confusing discourse relation pairs for RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$, respectively. Further details are included in Appendix B.1.

4.5.2 Experimental Results

Table 4.5 organizes the experimental results of second-level IDRR on the PDTB dataset. As the synthetic data focuses on learning infrequent discourse relations, it might cause the forgetting of frequent discourse relations and deteriorate Micro-F1. Despite the concern, we achieved the superior performance of both Micro-F1 and Macro-F1 in both RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$ thanks to the synthetic data.[7]

Detailed results are organized in Table 4.6. Regarding RoBERTa$_{\text{BASE}}$, the synthetic data is actually effective in learning infrequent discourse relations such as “Cause+Belief” and “Equivalence”. On the other hand, it does not work on

[7] Jiang et al. (2023) have also reported the performance of their RoBERTa$_{\text{LARGE}}$-based model is Micro-F1 of 66.4 and Macro-F1 of 60.1, which is the SOTA performance of RoBERTa$_{\text{LARGE}}$ to the best of our knowledge. However, we found their implementation of handling of multi-labeled examples might be different from ours, as mentioned in Section 4.3.1. Thus, we re-evaluated our model in their manner and confirmed the performance was Micro-F1 of 68.1 and Macro-F1 of 61.6, which still outperformed the SOTA performance.

Model	Setting		Micro-F1	Macro-F1
GPT-3.5	few-shot		23.2	19.0
GPT-4	few-shot		29.4	30.9
RoBERTa (BASE)	Vanilla		$64.2_{\pm 1.2}$	$57.1_{\pm 0.4}$
	Long and Webber (2022)		64.7	57.6
	Ours	+synthetic data (unfiltered)	$64.5_{\pm 0.8}$	$58.4_{\pm 1.2}$
		+synthetic data (vanilla-filtered)	$63.3_{\pm 0.5}$	$57.7_{\pm 0.6}$
		+synthetic data (LLM-filtered)	$\mathbf{64.8}_{\pm \mathbf{1.0}}$	$\mathbf{59.1}_{\pm \mathbf{1.5}}$
RoBERTa (LARGE)	Vanilla		$67.7_{\pm 0.5}$	$60.9_{\pm 1.6}$
	Ours	+synthetic data (unfiltered)	$67.6_{\pm 0.9}$	$62.1_{\pm 2.0}$
		+synthetic data (vanilla-filtered)	$67.9_{\pm 0.2}$	$62.0_{\pm 1.0}$
		+synthetic data (LLM-filtered)	$\mathbf{68.8}_{\pm \mathbf{0.4}}$	$\mathbf{62.4}_{\pm \mathbf{1.5}}$

Table 4.5: Experimental results of second-level IDRR on PDTB-3 dataset. The scores are the mean and standard deviation over three runs with different random seeds. The difference between synthetic data (unfiltered) and (filtered) is whether or not to apply the filtering by an LLM or a vanilla fine-tuned model in the second stage.

"Synchronous". One of the possible reasons is that the synthetic data may contain some phrases that induce shortcut learning, as discussed in Section 4.4.2. The above problem can be alleviated by refining the instruction so as not to include discourse connectives. Regarding RoBERTa$_{\text{LARGE}}$, the synthetic data is generally effective in learning the target discourse relations except "Cause+Belief". As the size of synthetic data for "Cause+Belief" is relatively small, the model may not have been adequately trained on the discourse relation.

Comparing the unfiltered and filtered synthetic data, we can see the solid performance gain thanks to the filtering by an LLM. We presume some removed examples are harmful to learning discourse relations even though they are not always noisy.

Relation	RoBERTa$_{\text{BASE}}$			RoBERTa$_{\text{LARGE}}$	
	VNL	Ours	L&W	VNL	Ours
Temporal.Synchronous	34.4	32.6♠	41.4	35.5	38.1♠
Temporal.Asynchronous	66.8	68.0	66.4	72.9	76.0
Contingency.Cause	69.3	69.8	71.4	74.1	74.8
Contingency.Cause+Belief	1.7	11.8♠	0.0	5.0	5.1♠
Contingency.Purpose	94.8	93.6	96.1	95.8	95.2
Contingency.Condition	70.2	73.8	74.1	75.5	78.0
Comparison.Concession	60.1	61.7	60.1	63.3	63.9
Comparison.Contrast	49.0	49.1	56.9	56.7	56.4
Expansion.Conjunction	60.6	60.0	61.7	62.9	65.0
Expansion.Equivalence	21.6	34.1♠	11.4	25.3	31.4♠
Expansion.Instantiation	69.8	72.7	69.8	73.1	73.2
Expansion.Level-of-detail	57.0	57.0	55.3	58.9	59.4
Expansion.Manner	80.3	80.5	78.4	80.9	83.1♠
Expansion.Substitution	63.7	62.3	63.8	72.7	73.9

Table 4.6: Detailed results of second-level IDRR on PDTB-3 dataset. VNL and L&W represent Vanilla and Long and Webber (2022), respectively. "Ours" corresponds to the "+synthetic data (LLM-filtered)" setting. ♠ denotes that the model was trained with synthetic data of the discourse relation.

4.5.3 Discussion

Effect of Top-*k* Table 4.7 shows the change in performance of RoBERTa$_{\text{BASE}}$ when varying how many confusing discourse relation pairs to extract. While synthetic data is generally effective in improving Macro-F1, learning to resolve more confusion does not necessarily lead to overall performance improvement, which suggests the importance of choosing which confusion to focus on.

Prompting from Discourse Connectives Inspired by previous studies, we attempted to generate synthetic data utilizing discourse connectives to examine their effectiveness. Let us explain based on Figure 4.6. We added an annotated

k	Micro-F1	Macro-F1
0	$64.2_{\pm 1.2}$	$57.1_{\pm 0.4}$
1	$63.6_{\pm 0.8}$	$57.5_{\pm 0.9}$
3	$\mathbf{64.8}_{\pm \mathbf{1.0}}$	$\mathbf{59.1}_{\pm \mathbf{1.5}}$
5	$63.9_{\pm 1.0}$	$58.4_{\pm 1.6}$

Table 4.7: Correspondence between the number of confusing discourse relation pairs to extract and the performance of RoBERTa$_{\text{BASE}}$.

discourse connective to the beginning of original Arg2 (i.e., "her political enemies ... " → "**because** her political enemies ... ") and made an LLM generate text that starts with the connective (i.e., "– {completion}" → "– **because** {completion}"). We incorporated the synthetic data generated by the above method and evaluated the model performance.

As a result, the performance of RoBERTa$_{\text{BASE}}$ was Micro-F1 of 64.2 and Macro-F1 of 58.7. The effect of discourse connectives on this setting is somewhat limited.

Single-Stage Augmentation Strategy We generated synthetic data by simply instructing an LLM to paraphrase argument pairs and investigated the performance gain by the synthetic data to compare with our strategy. In Figure 4.5, we modified the instruction to "Please write down paraphrases of ⟨pair of Arg1 and Arg2⟩ keeping the relation R_{true}" and obtained paraphrases of argument pairs. We incorporated the synthetic data and evaluated the model performance.

As a result, the performance of RoBERTa$_{\text{BASE}}$ was Micro-F1 of 64.4 and Macro-F1 of 57.3, which implies the importance of generating diverse Arg2.

4.6 Experiments on Japanese DRR

We also conducted experiments to investigate whether the proposed method is also effective in Japanese DRR.

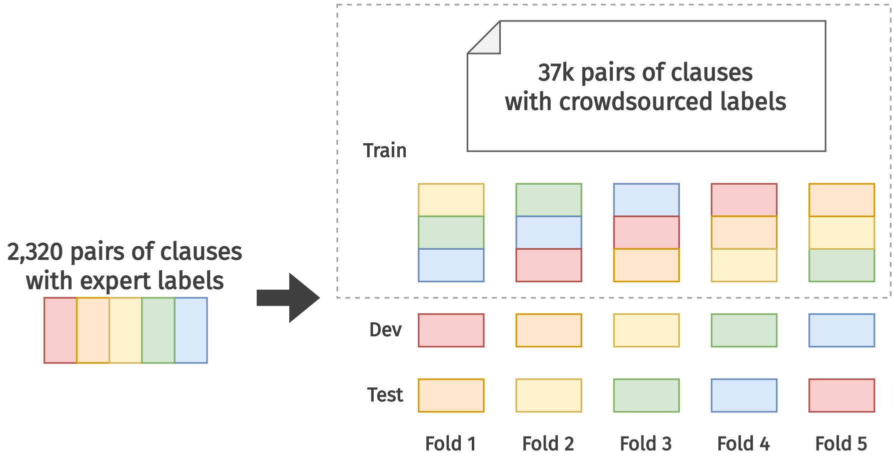


Figure 4.8: Illustration of five-fold cross-validation on KWDLC. We excluded some pairs of clauses from training data for each fold to ensure that there are no duplicate pairs of clauses between the training and development/test data. If some pairs of clauses in training data have both expert and crowdsourced labels, we prioritize the expert label.

4.6.1 Experimental Settings

Data and Model

We used the Kyoto University Web Document Leads Corpus (KWDLC)[8] (Kawahara et al., 2014; Kishimoto et al., 2018, 2020), which consists of the first three sentences of 6,445 web documents. All the documents have been annotated with discourse relations between clauses by crowdsourcing, and 500 out of 6,445 documents have also been annotated by linguistic experts. In this study, we evaluated the classification performance on 2,320 clause pairs with expert labels using five-fold cross-validation (cf. Figure 4.8). Unlike the experimental settings in Chapter 3, we utilized expert data that are not used for evaluation data for training because RoBERTa has achieved performance comparable to crowdworkers.

Synthetic data is generated from training data with expert labels for each fold. These statistics are organized in Table 4.8.

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

Relation	**Train**		**Synthetic Data**		**Dev**	**Test**
	Crowd	**Expert**	**Unfiltered**	**LLM-Filtered**		
Cause/Reason	2,350	149	505	425	47	46
Purpose	538	27	77	68	9	8
Condition	704	42	138	128	10	13
Justification	383	10	28	2	2	3
Contrast	380	5	-	-	0	1
Concession	759	44	150	142	30	26
No relation	31,264	1,110	-	-	380	377

Table 4.8: Statistics of the KWDLC dataset and synthetic data. For space limitation, we show the statistics of Fold 1 and the synthetic data for Japanese RoBERTa$_{\text{BASE}}$ as a representative.

Relation	**Definition**
Cause/Reason	one clause represents the cause or reason, while the other clause represents the result
Purpose	one clause represents the goal, while the other clause represents the means to achieve it
Condition	one clause represents the condition, while the other clause represents the result
Justification	one clause represents inference or recognition, while the other clause represents evidence
Contrast	emphasize the difference in the situations represented by both clauses
Concession	one clause negates the situation expressed by the other clause
No Relation	there is no semantic relation between clauses, or it represents a weak relation such as temporal, specification, instantiation, and so forth.

Table 4.9: Definitions of discourse relations in KWDLC. We refer to the KWDLC annotation manual (`https://github.com/ku-nlp/KWDLC/blob/master/doc/disc_guideline.pdf`).

The task is formulated as a seven-way classification of discourse relations between clauses, including “No Relation”. Table 4.9 organizes discourse relations defined in KWDLC. Unlike English IDRR, Japanese DRR does not distinguish between explicit and implicit discourse relations.

We evaluated the performance of the RoBERTa (Liu et al., 2019) model. For a pre-trained model, we employed the Japanese base-[9] and large-size[10] pre-trained

[9] `https://huggingface.co/nlp-waseda/roberta-base-japanese`

[10] `https://huggingface.co/nlp-waseda/roberta-large-japanese`

models hosted on Hugging Face Hub.

Training Details

Training details are almost the same as Section 4.5.1, but we incorporate crowdsourced data into training. Thus, the objective function is as follows:

$$H = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{f_y(x)}}{\sum_{y' \in [Y]} e^{f_{y'}(x)}}$$

$$L = H_{\text{expert}} + \lambda_c \times H_{\text{crowd}} + \lambda_s \times H_{\text{synthetic}}$$

where λ_c is the weight for crowdsourced data.

During the evaluation phase, we evaluate the model by Micro-F1 and Macro-F1. We measure the performance on the development split per epoch and adopt the model parameters with the best dev Micro-F1[11] for evaluation on the test split.

Hyper-Parameters

Regarding baselines, we performed a grid search of learning rate from {1e-5, 2e-5} and chose the one that achieved the best Micro-F1 on the development split. When incorporating crowdsourced and synthetic data into training, we used the same hyper-parameters but performed grid search of λ_c and λ_s, from {0.5, 0.25}. As we generated three synthetic data from top-1, 3, and 5 confusing discourse relation pairs, we adopted the one that achieved the best Micro-F1 on the development split. Specifically, we used the synthetic data generated from top-5 confusing discourse relation pairs for both RoBERTa$_{\text{BASE}}$ and RoBERTa$_{\text{LARGE}}$. Further details are included in Appendix B.1.

4.6.2 Experimental Results

Table 4.10 shows the experimental results of DRR on KWDLC. It is observable that synthetic data is also effective in Japanese DRR.

[11] We focus on Micro-F1 as the number of expert labels for some discourse relations is quite small, and thus Macro-F1 is somewhat unstable.

Model	Setting		Micro-F1	Macro-F1
RoBERTa (BASE)	Vanilla		$51.6_{\pm1.3}$	$42.7_{\pm1.7}$
	Ours	+synthetic data (unfiltered)	$53.0_{\pm1.6}$	$\mathbf{44.0_{\pm2.0}}$
		+synthetic data (vanilla-filtered)	$51.5_{\pm1.6}$	$42.3_{\pm1.8}$
		+synthetic data (LLM-filtered)	$\mathbf{53.3_{\pm0.3}}$	$43.6_{\pm1.0}$
RoBERTa (LARGE)	Vanilla		$52.7_{\pm0.5}$	$41.6_{\pm1.0}$
	Ours	+synthetic data (unfiltered)	$53.0_{\pm0.7}$	$43.8_{\pm1.1}$
		+synthetic data (vanilla-filtered)	$51.8_{\pm1.3}$	$43.2_{\pm0.8}$
		+synthetic data (LLM-filtered)	$\mathbf{53.2_{\pm1.7}}$	$\mathbf{44.9_{\pm1.7}}$
Human (crowdworker) (Kishimoto et al., 2020)			51.5	**46.0**

Table 4.10: Experimental results of DRR on KWDLC. The scores are the mean and standard deviation over three runs with different random seeds.

Compared with crowdworkers, there is room for improvement in Macro-F1. One possible reason could be that the number of some discourse relations is too small to learn, even with synthetic data.

4.7 Summary of This Chapter

We proposed a method of generating synthetic data for DRR using an LLM, which is summarized as two steps: extraction of confusing discourse relation pairs based on false negative rate and generation of synthetic data focused on resolving the confusion. According to the proposed method, we built synthetic data effective in DRR while addressing the complexity of DRR by two-stage prompting. As a result of experiments, we demonstrated its effectiveness both in English and Japanese DRR.

One of the future directions is to explore the collaboration of an encoder-only language model and an LLM for reasoning-aware discourse relation recognition. We would also like to consider a method to generate synthetic data from scratch for further scalability.

Chapter 5

Application to Japanese Writing Education

5.1 Introduction

In this chapter, we introduce an educational application utilizing the data constructed in the process of our studies, taking a slight departure from previous chapters. We aim to demonstrate the usefulness of discourse relations through the implementation of the application.

As we use natural language as a means of cognition and communication, language education plays an indispensable role in our lives. In language education, written language production is also crucial for learning how to express our thoughts.

However, in Japanese writing education, it is a long-standing problem that elementary school students tend to have an aversion to writing compositions (National Institute for Educational Policy Research, 2008; Ritsumeikan University Library, 2017). One of the possible reasons is that they learn Japanese writing only through open-ended writing assignments such as book reports and essays on topics related to daily life. Such assignments rarely motivate the students to go further than merely complete them, and they often struggle to come up with what to write due to the open-ended format. In addition, they receive little feedback on their writing, leading to a vicious cycle where they become increasingly aware of being poor at writing without learning how to improve.

In order to ameliorate the current situation, Japanese education is in need of educational material that offers a more engaging experience, i.e., enables students to construct sentences with fun and get feedback on their writing. However, there are two major challenges to achieving this: how to reduce an aversion to constructing sentences and how to evaluate writing automatically.

Game-based learning (Vandercruysse et al., 2012), which aims to make the learning process more fun with a game, is a promising solution to the first challenge. There has been no AI educational game for studying Japanese writing to the best of our knowledge; thus, it is worth attempting to design such a game and investigate its effectiveness.

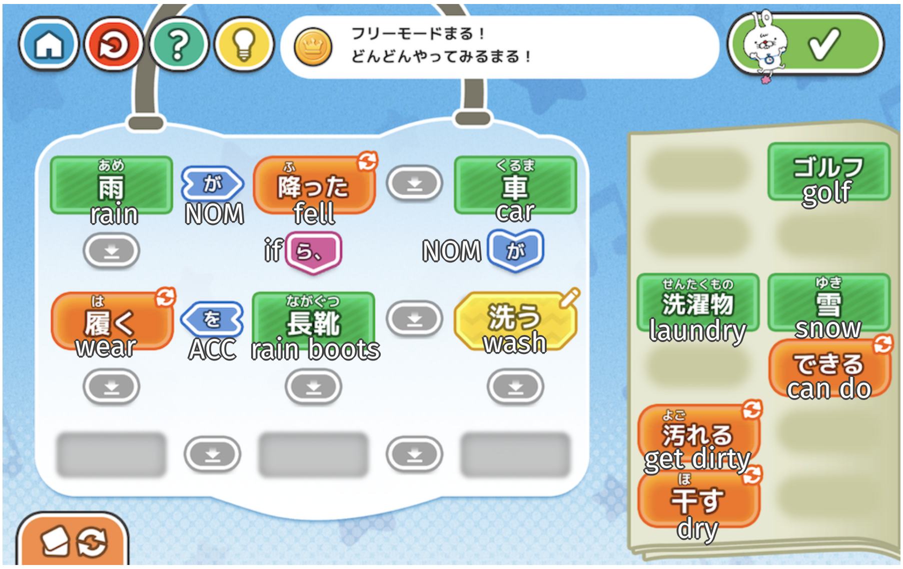
Regarding the second challenge, we focus on existing language resources. In natural language processing (NLP), large-scale language resources have been built so far to teach linguistic and world knowledge to computers (Fellbaum, 1998; Baker et al., 1998; Kawahara and Kurohashi, 2006; Speer et al., 2017; Sap et al., 2019), some of which can be utilized for human learning as well as machine learning. By exploiting them, it is now possible to automatically generate and score simple writing practice questions, and the aforementioned educational material has become more feasible.

This study aims to develop an educational game for elementary school students to study Japanese writing with fun and investigate its effectiveness. To these ends, we design an AI educational game utilizing existing language resources. Hereafter, we call it "Kotoba-musubi".[1]

Kotoba-musubi is a word-based game where the player builds simple sentences by connecting content words with case markers and combining the simple sentences with discourse markers into complex sentences with contingent relations (Figure 5.1). Players apply given word cards to empty rectangle frames and connect them using arrow-shaped particle marks.

The created sentences are automatically scored using large-scale language resources, and players can obtain feedback on the spot. For instance, if no examples

[1] "Kotoba" and "musubi" are the Japanese words that mean "word" and "connecting" in English, respectively.



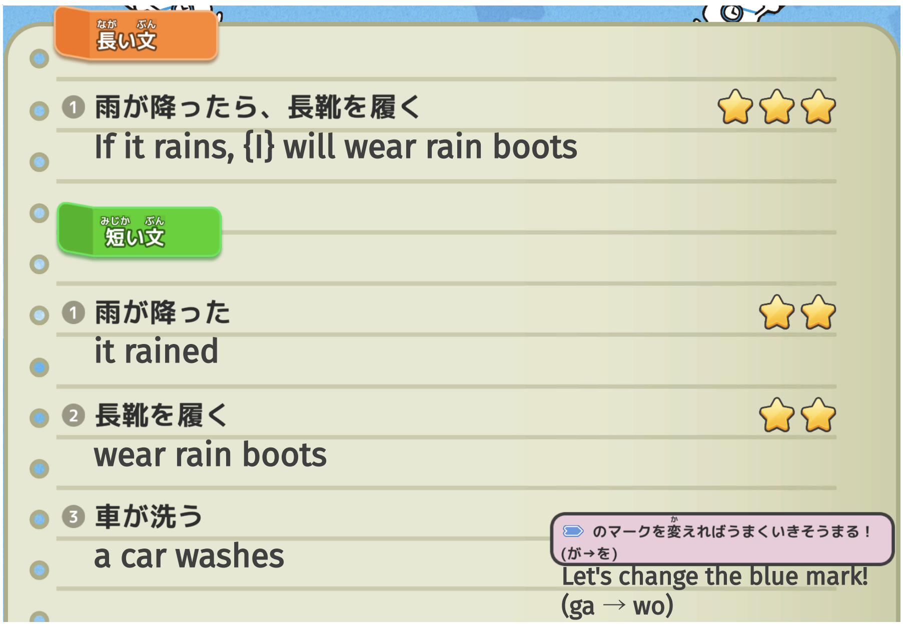
フリーモードまる！
どんどんやってみるまる！

Figure 5.1: Screenshot of the play screen of Kotoba-musubi with gloss. Kotoba-musubi is a word-based game, where the player builds simple sentences and complex sentences with contingent relations by connecting given word cards with particle marks. "NOM" and "ACC" in the figure stand for nominative and accusative cases, respectively.

of a simple sentence are found in the Japanese case frames (Kawahara and Kurohashi, 2006) due to incorrect usage of a case marker, our system gives feedback suggesting a more appropriate case marker, as illustrated in Figure 5.2. Players can re-arrange their compositions and retry the automatic scoring; thus, our system has interactivity that writing assignments do not have.

Kotoba-musubi focuses on learning how to construct basic sentences with fun, and it is intended to be used as an introduction to studying Japanese writing. We expect that, through playing the game, students discover the enjoyment of considering sentences and cultivate knowledge necessary for writing, such as collocation, usage of words and particles, and contingent relations between basic event expressions.

We also develop smartphone and web applications of Kotoba-musubi and conduct a user study to investigate its effectiveness. We have released Kotoba-musubi



長い文

① 雨が降ったら、長靴を履く ★★★
If it rains, {I} will wear rain boots

短い文

① 雨が降った ★★
it rained

② 長靴を履く ★★
wear rain boots

③ 車が洗う
a car washes

🟦 のマークを変えればうまくいきそうまる！（が→を）
Let's change the blue mark!
(ga → wo)

Figure 5.2: Scoring results of the sentences in Figure 5.1. {} indicates a dropped pronoun.

for educational purposes, which is available at `https://nlp.ist.i.kyoto-u.ac.jp/EN/?Kotobamusubi`.

The contributions of this study are summarized as follows:

- We propose Kotoba-musubi, an educational game for elementary school students to study Japanese writing utilizing existing language resources and AI.
- We developed smartphone and web applications of Kotoba-musubi and conducted a user study to assess engagement.
- As a result of the user study, we demonstrated our game can be used as a good introduction to studying Japanese writing.

5.2 Related Work

Word games similar to Kotoba-musubi, where the players connect smaller linguistic units to build larger ones, have been developed in the past. However, as represented by Scrabble[2], most of them are character-based, where the players build words by connecting characters. It is relatively easy to implement such games because these games need only a vocabulary dictionary to judge whether or not words have been produced successfully. In contrast, these games cannot handle sentences, which limits the knowledge that can be acquired from them.

Game-based learning and gamification have been actively studied to make the learning process more fun. While the former refers to learning that makes use of educational games with defined learning outcomes (Vandercruysse et al., 2012), the latter refers to the application of game elements to non-game problems (Deterding et al., 2011). Kotoba-musubi is developed for game-based learning and is an AI educational game aiming at making constructing sentences more fun. Although several studies have proposed game-based learning with AI (Dyulicheva and Glazieva, 2021), there has existed no AI educational game for studying Japanese writing to the best of our knowledge. Regarding gamification, Duolingo[3] is one of the educational applications that use gamification for learning language, including written language production. While not an educational game, it introduces game elements such as rewards and a badge collection feature to motivate learners.

5.3 Kotoba-musubi

5.3.1 Game Design

In a single round, players are given 12 word cards and nine particle marks. The objective is to achieve a higher score by constructing basic simple sentences, and complex sentences with contingent relations using the cards and marks. In order to construct sentences, players arrange the given cards/marks according to the empty rectangle/oval frames on the screen. The score of each composition is automatically computed by pattern matching with language resources.

[2] https://en.wikipedia.org/wiki/Scrabble

[3] https://ja.duolingo.com/

Word Card The 12 word cards break down into six noun cards, five verb/adjective cards, and one wildcard that allows players to enter a word freely.[4] The role of each card is distinguishable by its color; therefore, players can learn how to construct sentences even if they do not understand the concept of parts of speech. The word cards can also be distinguished by shape, considering those with color vision deficiency. In addition, ruby characters are written above Chinese characters to facilitate reading. Regarding verb/adjective cards, the predicates can be conjugated in the present, past, negative, or past-negative form, which makes it possible to create more diverse sentences.

Particle Mark The nine particle marks break down into five major case particles and four discourse connectives representing contingent relation. The five marks consist of "が [ga]", "を [wo]", "に [ni]", "で [de]", and "と [to]", which roughly correspond to nominative case, accusative case, dative case, instrumental case, and "with" or "and", respectively. The four marks consist of "から [kara]", "ので [node]", "と [to]", and "ら [ra]", the first and last two of which represent causal and conditional relations, respectively. As with the word cards, the role of each mark is distinguishable by its color and shape. Each mark is arrow-shaped, the direction of which indicates the dependency between words.

Referring to Figure 5.1, the player arranges the word cards "雨 (rain)" and " 降った (fell)" next to each other and connects them using the particle mark "が (nominative case)". The direction of the mark is from left to right; therefore, the player has built the simple sentence "雨が降った (it rained)". Furthermore, it is connected to the simple sentence "長靴を履く (wear rain boots)" with the particle mark "ら (if)"; that is, the player has also constructed the complex sentence "雨が降ったら, 長靴を履く (If it rains, I[5] will wear rain boots)".

[4] The input is automatically analyzed by the Japanese morphological analyzer, Juman++ (Morita et al., 2015; Tolmachev et al., 2018).

[5] {} indicates a dropped pronoun.

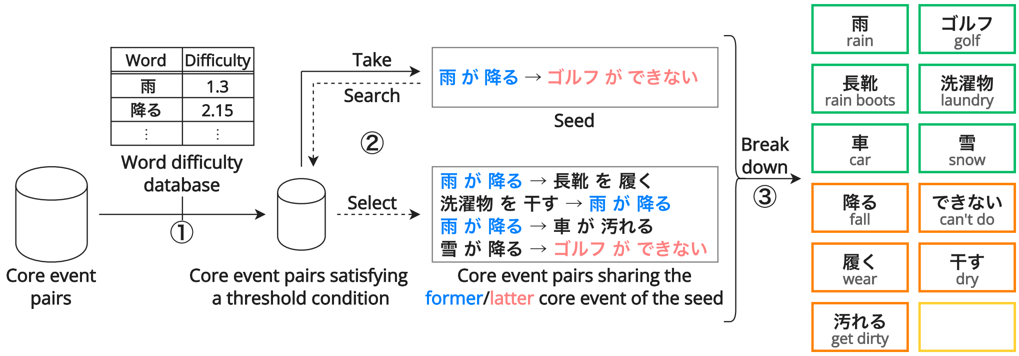

Word
Difficulty
雨
1.3
降る
2.15
Word difficulty database
Core event pairs
①
Take
Search
②
Select
Core event pairs satisfying a threshold condition
雨 が 降る → ゴルフ が できない
Seed
雨 が 降る → 長靴 を 履く
洗濯物 を 干す → 雨 が 降る
雨 が 降る → 車 が 汚れる
雪 が 降る → ゴルフ が できない
Core event pairs sharing the former/latter core event of the seed
Break down
③
雨 rain
ゴルフ golf
長靴 rain boots
洗濯物 laundry
車 car
雪 snow
降る fall
できない can't do
履く wear
干す dry
汚れる get dirty

Figure 5.3: Overview of the method for generating word card sets.

5.3.2 Method for Generating Word Card Sets

In order for the game to be more fun, it is preferable that players can construct several simple and complex sentences from a given word card set. To guarantee this, we focus on **core event**, which is defined in Section 2.3.1.

We adopt pairs of core events that have contingent relation (**core event pairs**) such as "雨が降る → 長靴を履く (it rains → wear rain boots)", which can be extracted at scale from the Kyoto University Commonsense Inference dataset, the dataset constructed in Chapter 2.

The proposed method is to automatically generate word card sets from core event pairs, which consists of the following three steps (cf. Figure 5.3).

1. Set a threshold regarding word difficulty and exclude core event pairs not satisfying the condition.

2. Take one core event pair as a **seed** and randomly select the other four pairs that share the former or latter core event of the seed.

3. Generate a word card set by breaking down the selected five pairs into predicates and arguments.

STEP 1: Filtering by Word Difficulty First, we adjust vocabulary using the Japanese word difficulty database (Muzitani et al., 2019), considering that the target users are elementary school students. This database contains 26k words

Number	Acquisition Time
1	Before elementary school
2	Early elementary school
3	Late elementary school
4	After junior high school
5	Never seen or heard

Table 5.1: Correspondence between numbers and acquisition times in the Japanese word difficulty database (Muzitani et al., 2019).

of basic Japanese vocabulary labeled with their average acquisition time using crowdsourcing. Acquisition time is regarded as word difficulty and expressed as a number from 1 to 5, the correspondence of which is shown in Table 5.1.

In this study, we set the following threshold conditions.

Easy (for early elementary school students) The maximum word difficulty of words in a core event pair does not exceed 2.0.

Medium (for middle elementary school students) The average word difficulty of words in a core event pair exceeds 1.5, and the maximum word difficulty does not exceed 2.5.

Hard (for late elementary school students) The average word difficulty of words in a core event pair exceeds 2.0.

We set a strict upper bound but allow easy words to get mixed in harder sets.

STEP2: Selecting Core Event Pairs Then, we select five core event pairs from the ones satisfying a threshold condition. Specifically, we take one core event pair (hereafter, **seed**) and randomly select the other four pairs that share the former or latter core event of the seed. If we fail to get five pairs, including a seed, we skip generating a word card set.

STEP3: Generating a Word Card Set Finally, we generate a word card set by breaking down the five pairs selected in the previous step into predicates and arguments. If we obtain six nouns and five or six verbs/adjectives after deduplication, we regard the words as a word card set. In case six verbs/adjectives are obtained, we replace the least frequent verb/adjective with a wildcard.

5.3.3 Careful Examination of Core Event Pairs

We must be extremely careful not to give inappropriate questions as an educational application. Accordingly, we request two linguistic experts to manually classify the core event pairs used for word card sets into the following categories.

Valid A core event pair has contingent relation.

Invalid A core event pair has no contingent relation.

Inappropriate A core event pair contains educationally inappropriate expressions.

Prior to the examination, we exclude core event pairs that contain minor case particles or words unregistered in the Japanese word difficulty database. As a result, 9k core event pairs are left and examined. We also automatically assign reading to each word and conjugated forms to each predicate using the Japanese morphological analyzer, Juman++ (Morita et al., 2015; Tolmachev et al., 2018). We ask the experts to correct auto-assigned readings in addition to the examination.

As a result of the examination by the experts, we finally obtained 4,362 valid, 3,560 invalid, and 1,120 inappropriate core event pairs, respectively. We used the "valid" core event pairs for generating word card sets.

5.3.4 Statistics of Word Card Sets

According to the method described in Section 5.3.2, we generated word card sets from the verified pairs. As a result, we obtained 99 easy, 594 medium, and 284 hard word card sets, respectively, achieving a size that is sufficient for solo play.

5.3.5 Automatic Scoring

When players tap a scoring button, the board information is sent to a back-end server and automatically scored. The automatic scoring is performed in the following three steps:

1. Recognize simple and complex sentences.
2. Score each sentence automatically.
3. Generate feedback based on the scoring results.

Recognition of Simple and Complex Sentences

A simple sentence is recognized as a sequence of a verb card at the end and noun cards in the rest connected by case particle marks, and a complex sentence as two simple sentences connected by a discourse connective mark. We consider all possible combinations of simple sentences.

Automatic Scoring of Each Sentence

Our policy is to prioritize simple sentences whose predicate and argument frequently co-occur (i.e., idiomatic) or whose length is longer. The score of each simple sentence is determined based on the number of examples in the Japanese case frames (Kawahara and Kurohashi, 2006). Specifically, regarding the case frame *cf* where examples are found, we compute a score of each argument and case pair *(a, c)* with the following function S and sum it up.

$$S(a, c, cf) = 0.5 + 0.5 \times min(1, \frac{f_{a,c,cf}}{f_{cf}} \times 5)$$

where $f_{a,c,cf}$ is the frequency of argument a in case c of case frame *cf*, and f_{cf} is the frequency of case frame *cf*. For instance, referring to Table 5.2, the score of the simple sentence “筍が顔を出す (Bamboo shoots come out from the ground)” is computed as follows.

$$\frac{f_{\text{筍, が, 出す}_2}}{f_{\text{出す}_2}} = \frac{199}{605,564} = 0.000, \qquad \frac{f_{\text{顔, を, 出す}_2}}{f_{\text{出す}_2}} = \frac{256,404}{605,564} = 0.423$$

$$S(\text{筍, が, 出す}_2) + S(\text{顔, を, 出す}_2) = 0.5 + 1 = 1.5$$

Case slots	**Case fillers**
が [ga] $_{63,664}$	太陽 (sun) $_{6,481}$, ... , 筍 (bamboo shoot) $_{199}$, ...
を [wo] $_{320,338}$	顔 (face) $_{256,404}$, 口 (mouth) $_{41,127}$, ...
に [ni] $_{64,750}$	店 (shop) $_{2,154}$, 実家 (parents' house) $_{1,632}$, ...
$f_{出す_2}$: 605,564	

Table 5.2: Second case frame of the verb "出す (show)". "が [ga]", "を [wo]", and "に [ni]" roughly correspond to nominative, accusative, and dative cases, respectively. The number following a case or a case filler represents its frequency.

Each score is converted into the following symbols for further interpretability.

$$r = \text{☆☆ (score} >= 1), \text{☆} \ (0 < \text{score} < 1), ?(\text{score} = 0)$$

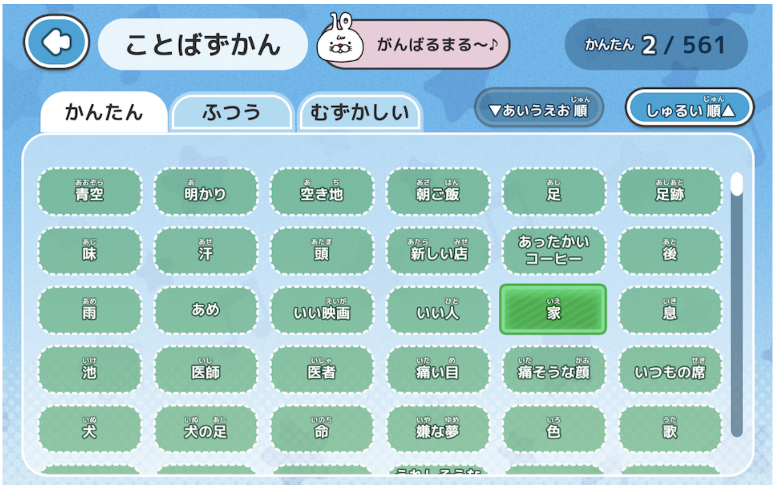
Regarding complex sentences, they are graded as "☆☆☆" if they contain one of the verified core event pairs; otherwise, as "?". When examining the containment relation, we take into account the polarity of the negation of a predicate.

Generation of Feedback based on the Scoring Results

The following feedback is generated from sentences graded as "?" depending on the reasons behind the sign's attribution.

- There is a more appropriate case marker that allows for the sentence to achieve a score.
- The sentence is grammatically incorrect (e.g., the sentence starts with a verb/adjective, a verb/adjective depends on a noun, and so forth).
- The sentence is not matched with any example in language resources.

Regarding the third point, we expect to collect unknown contingent relations through an error reporting function and thus improve the evaluation system.



青空	明かり	空き地	朝ご飯	足	足跡
味	汗	頭	新しい店	あったかいコーヒー	後
雨	あめ	いい映画	いい人	家	息
池	医師	医者	痛い目	痛そうな顔	いつもの席
犬	犬の足	命	嫌な夢	色	歌

Figure 5.4: Screenshot of "word dictionary" feature in Kotoba-musubi.

5.3.6 Collection Element

We introduce the "word dictionary" feature to motivate students to continue to play the game (Figure 5.4). It records which words each player has used so far in the game by their difficulty level. Regarding each word in the word dictionary, we can also confirm its reading and refer to example sentences that contain the word regardless of whether or not each player has ever used it. We believe this feature prompts users to accept further challenges and thus to enlarge their word collection and develop their vocabulary.

5.3.7 Implementation

Kotoba-musubi is a client-server application that is available on iOS, Android, and web browsers. The client side is developed with Unity and runs on modern web browsers using WebGL. The server side consists of a Nginx web server and an application server developed with the Python web framework, Flask. The application server has 96GB RAM and 24 cores. The account sign-in function is implemented using Firebase Authentication.

5.4 User Study

We developed the iOS, Android, and web applications that implement Kotoba-musubi and conducted a user study to assess engagement.

5.4.1 Settings

We recruit 80 pairs of elementary school students and their parents across Japan and have the children play the game for an hour in total over two days. 80 students consist of 10 boys and 10 girls each from third to sixth grade in elementary school. In order to let them play freely, we neither set their quota nor specify the difficulty of questions they tackle. How to play is displayed in the first play and can be confirmed at any time. After playing the game, they answer the questionnaire described in Table 5.3.

5.4.2 Results

Figure 5.5 illustrates the aggregate results of the questionnaire. We can see that elementary school students tend to dislike writing compositions, as mentioned in Section 5.1. Despite this disadvantageous situation, 70% of the participating children enjoyed Kotoba-musubi, and 90% answered it was worth playing the game. In addition, 70% expressed their will to continue to play the game, which suggests that it is a good introduction to studying Japanese writing.

We also investigated the number of children who disliked writing compositions but enjoyed the game. We found that 18 of 37 children who disliked or disliked a little writing compositions enjoyed the game[6]. This result supports the effectiveness in reducing an aversion to writing compositions.

Table 5.4 lists the excerpted feedback comments from the participating children. While a number of children noted they enjoyed themselves, there were also several comments that it took time to understand how to play, which raises the need for improving a tutorial.

[6] Specifically, 10, 8, 10, 7, and 2 children answered “Enjoyed”, “Enjoyed a little”, “Neither”, “Didn’t enjoy a little”, and “Didn’t enjoy”, respectively.

Do you like writing compositions?				
5. Like	4. Like a little	3. Neither	2. Dislike a little	1. Dislike
Did you enjoy playing Kotoba-musubi?				
5. Enjoyed	4. Enjoyed a little	3. Neither	2. Didn't enjoy a little	1. Didn't enjoy
Do you want to continue to play Kotoba-musubi?				
5. Yes	4. Yes, a little	3. Neither	2. No, a little	1. No
Which do you think Kotoba-musubi is: "game" or "study"?				
5. Game	4. A little game	3. Neither	2. A little study	1. Study
Were you unsatisfied with the automatic scoring?				
5. No	4. No, a little	3. Neither	2. Yes, a little	1. Yes
Do you think it was worth playing Kotoba-musubi?				
2. Yes	1. No			

Table 5.3: Main items and options of the questionnaire answered by the participating children. Regarding the second and subsequent items, the option with a larger number is more preferable.

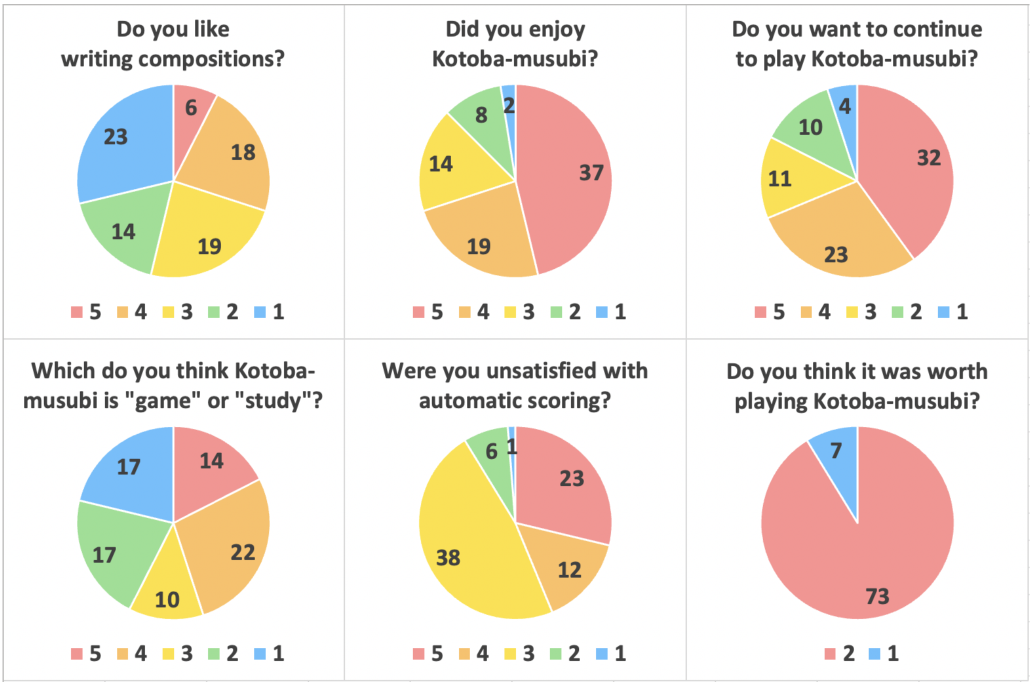


Figure 5.5: Aggregate results of the questionnaire described in Table 5.3. The numbers in the figure represent those of children who chose the option.

+ It was fun and refreshing to create sentences using words I didn't usually use.
+ I'd honestly like to see the game introduced to school tablets.
± It will be more fun if we can compete with our friends for higher scores.
− I couldn't fully understand how to play the game.
− It was hard to get sentences scored, which made me frustrated.

Table 5.4: List of the excerpted feedback comments from the participating children.

5.4.3 Discussion

The user study also revealed some current issues of Kotoba-musubi. For instance, there is room for improvement in reducing the feeling of studying while playing the game. As one of the feedback comments in Table 5.4 suggests, we need to enhance the enjoyment of our application by introducing game elements such as a match game system.

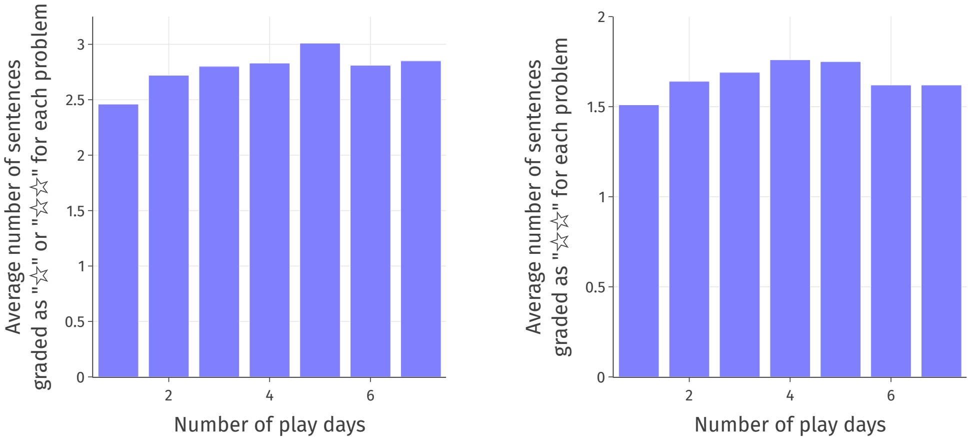


Figure 5.6: Analysis results of user logs.

Another issue is the quality of automatic scoring. Although the result of the fifth question in Table 5.5 leans toward positive, about half of the participants are also ambivalent. One of the possible remedies is to incorporate masked language models such as BERT (Devlin et al., 2019) into automatic scoring for more flexible evaluation. For instance, we can quantify to some extent the correctness of the usage of a case marker by masking it and predicting the probability of the original token in a masked position. We will explore how to utilize neural language models considering the computational cost.

5.4.4 Analysis of User Logs

We investigated whether there was some improvement as the number of play days increased. Specifically, we targeted 158 users with at least one week's worth of activity logs, aggregated the activity logs on a daily basis, and calculated the average number of sentences graded as "☆" or "☆☆" for each problem. We assume that high-frequency predicate-argument structures in case frames (i.e., sentences graded as "☆☆") are high quality.

Figure 5.6 illustrates the analysis results. For instance, as of day one and five, the average numbers of sentences graded "☆" or "☆☆" for each problem are 2.46 and 3.01. It is observable that users created more scored sentences as the number of play days increased, though there was some decrease after day six possibly because they became less motivated. In addition, we can see that the average number of sentences graded as "☆☆" gradually increased until day 5. This implies the effectiveness of our game in learning Japanese writing.

5.5 Summary of This Chapter

We proposed Kotoba-musubi, an AI educational game for elementary school students to study Japanese writing, which fully utilizes existing language resources such as case frames, a word difficulty database, and a commonsense contingency reasoning dataset. While playing the game, players construct simple and complex sentences by connecting given word cards with particle marks. We expect students to develop their vocabulary and reasoning skills in a ludic manner with the game.

We also developed smartphone and web applications of Kotoba-musubi and conducted a user study involving 80 pairs of elementary school students and their parents to assess it. The results of the user study demonstrated its effectiveness in reducing an aversion to writing compositions.

One future work is to address the remaining issues with reference to the feedback comments and further investigate long-term educational effects. We also consider collecting unknown contingent relations through error reports.

Chapter 6

Conclusion

6.1 Overview

The objectives of this thesis are two folds: to improve the linguistic capability to infer discourse relations primarily in Japanese and to explore its applications to other NLU tasks and human learning. Toward these objectives, we have studied data generation approaches.

In Chapter 2, we proposed a method of semi-automatically generating multiple-choice questions that ask basic contingency from a raw corpus. According to the proposed method, we built a large-scale Japanese commonsense contingency reasoning dataset comprising 104k problems. We demonstrated there was a reasonable performance gap between the NLP model at the time and humans on this dataset.

In Chapter 3, we worked on improving the linguistic capability to infer basic contingency utilizing the constructed dataset. We automatically generated large-scale pseudo-problems by utilizing the scalability of the aforementioned proposed method and improved commonsense contingency reasoning by incorporating them into training. We also investigated the generality of knowledge about basic contingency through quantitative evaluation by performing transfer learning from a commonsense contingency reasoning task to the related NLU tasks. Through these experiments, we confirmed the performance gain in both a commonsense contingent reasoning task and the related NLU tasks and thus demonstrated the importance of contingency reasoning in NLU.

In Chapter 4, we worked on improving Discourse Relation Recognition (DRR) with a synthetic data generation approach. Regarding DRR, one of the biggest issues is the paucity of training data for some error-prone discourse relations. To alleviate this issue, we proposed a method of generating synthetic data for these error-prone discourse relations using a large language model. Thanks to the synthetic data generated according to the proposed method, we achieved the performance gain in both Japanese and English DRR.

In Chapter 5, we introduced an educational application utilizing the data constructed in the process of our studies. We took up the long-standing problem in Japanese education that elementary school students tend to have an aversion to writing compositions and developed an educational game in order to ameliorate the situation. As the result of a user study, we demonstrated that our game can be used as a good introduction to studying Japanese writing. This also supports the usefulness of discourse relations.

6.2 Future Prospects

6.2.1 Evaluation of the Linguistic Capability to Infer Discourse Relations in Multi-Modal Settings

In this thesis, we evaluated/improved the linguistic capability of computers to infer discourse relations in a single-modal setting. However, with the rapid advancement of multi-modal models, it is becoming increasingly necessary to evaluate the extent to which they can infer discourse relations and act accordingly. Let us consider assistant robots that assist you with housework. If you instruct them to do laundry, and they can find some laundry should not be machine washed, i.e., understand contingent relation between “laundry” and “machine wash”, they can prevent the incident. For another instance, if you ask a robot to fetch a book without knowing that its hands are dirty, the robot should clean its hand before granting the request based on knowledge about basic contingency. Thus, the linguistic capability to infer discourse relations is crucial for robots that can function in our world. Toward such a thoughtful multi-modal model, it is deemed necessary to implement the evaluation framework of Discourse Relation Recognition in multi-modal settings.

There might be a claim that such discourse relations can also be learned solely from visual information. However, it is probably challenging and inefficient to learn vast knowledge about discourse relations from scratch based on visual information. We hope that our studies can contribute to this issue.

6.2.2 Exploration of Other Practical Applications

We explored the two applications of discourse relations in this thesis: applications of basic contingency to improving other NLU tasks and to Japanese writing education. Discourse relations are also expected to be useful for challenges in our world such as analyzing the stock market and identifying the pros and cons of a product or service, as described in Section 1. It is beneficial to explore such other practical applications for the public welfare.

One point to be considered for such applications is that our studies have addressed Discourse Relation Recognition in web or news domain. It is still under-explored whether computers can recognize discourse relations in text from other domains, such as financial and medical texts. There will be a need to focus on the generality of Discourse Relation Recognition in the future.

6.2.3 Rethinking of Evaluation of Natural Language Understanding

In this thesis, we took a bottom-up approach to NLU, i.e., worked on evaluating/improving the linguistic capability to infer discourse relations alone and did not work on defining language understanding. However, several studies have claimed the need to define language understanding anew and the importance of multifaced evaluation of NLU (Bender and Koller, 2020; Bommasani et al., 2021). In light of this trend, we should face this issue and pursue how natural language should be evaluated toward the ultimate goal of NLP.

Appendix A

Supplementary Materials of Chapter 3

A.1 Hyper-parameters

Table A.1, A.2, A.3, A.4, and A.5 organize the hyper-parameters used in the experiments. We found that a 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	XLM-R
Epoch	3	
Batch size	32	
Max sequence length	128	
Optimizer	AdamW	
Learning rate	2e-5	5e-6
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.1	
Seed	{0, 1, 2}	

Table A.1: Hyper-parameters for fine-tuning on KUCI with pseudo-problems.

Name	Value	
	BERT	XLM-R
Epoch	100	
Batch size	256	
Max sequence length	128	
Optimizer	AdamW	
Learning rate	1e-4	
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.06	
gradient clipping value	-	0.25
Seed	0	

Table A.2: Hyper-parameters for AMLM. Most of the hyper-parameters are referred to Gururangan et al. (2020).

Name	Value	
	BERT	XLM-R
Epoch	10	
Patience for early stopping	3	
Batch size	32	
Max sequence length	128	
Optimizer	AdamW	
Learning rate	2e-5	5e-6
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.1	
Seed	{0, 1, 2}	

Table A.3: 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
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.1	
Seed	{0, 1, 2}	

Table A.4: Hyper-parameters for fine-tuning on JWSC. We set the number of epochs to a large value with reference 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	
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.1	
Seed	{0, 1, 2}	

Table A.5: Hyper-parameters for fine-tuning on JCQA.

Appendix B

Supplementary Materials of Chapter 4

B.1 Hyper-parameters

Table B.1 and B.2 organize the hyper-parameters used in the experiments.

Name	Value	
	RoBERTa$_{\text{BASE}}$	RoBERTa$_{\text{LARGE}}$
Epoch	20	
Batch size	32	
Max sequence length	128	
Optimizer	AdamW	
Learning rate	2e-5	1e-5
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.1	
Seed	{0, 1, 2}	
top-k	3	5
λ	0.25	

Table B.1: Hyper-parameters regarding IDRR on the PDTB dataset.

Name	Value	
	RoBERTa$_{\text{BASE}}$	RoBERTa$_{\text{LARGE}}$
Epoch	100	
Batch size	256	
Max sequence length	128	
Optimizer	AdamW	
Learning rate	2e-5	1e-5
Scheduler	Linear decay with linear warmup	
Warmup proportion	0.06	
gradient clipping value	1.0	0.5
Seed	{0, 1, 2}	
top-k	5	
λ_c	0.25	0.5
λ_s	0.5	0.25

Table B.2: Hyper-parameters regarding DRR on KWDLC.

Bibliography

[1] Eugene Agichtein and Luis Gravano. Snowball: Extracting Relations from Large Plain-Text Collections. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, page 85–94, New York, NY, USA, 2000. Association for Computing Machinery. doi: 10.1145/336597.336644.

[2] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining Neural Scaling Laws, 2021.

[3] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet Project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August 1998. Association for Computational Linguistics. doi: 10.3115/980845.980860.

[4] Madeleine Bates. Models of natural language understanding. *Proceedings of the National Academy of Sciences*, 92(22):9977–9982, 1995. doi: 10.1073/pnas.92.22.9977.

[5] Robert De Beaugrande and Wolfgang U. Dressler. *Introduction to Text Linguistics*. Longman, London, 1981.

[6] Emily M. Bender and Alexander Koller. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463.

[7] Prajjwal Bhargava and Vincent Ng. DiscoSense: Commonsense Reasoning with Discourse Connectives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10295–10310, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.703.

[8] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439. Association for the Advancement of Artificial Intelligence, Apr. 2020. doi: 10.1609/aaai.v34i05.6239.

[9] Daniel G. Bobrow. Natural Language Input for a Computer Problem Solving System. Technical report, USA, 1964.

[10] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, ..., and Rohith Kuditipudi. On the Opportunities and Risks of Foundation Models, 2021.

[11] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075.

[12] 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, ..., and Dario Amodei. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[13] Lynn Carlson and Daniel Marcu. Discourse Tagging Reference Manual, 2001. URL `https://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf`.

[14] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurovsky. Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001.

[15] Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. RST Discourse Treebank, 2002.

[16] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001.

[17] Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Wong, and Simon See. DiscoPrompt: Path Prediction Prompt Tuning for Implicit Discourse Relation Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 35–57, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.4.

[18] Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45, 2003.

[19] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the*

Association for Computational Linguistics, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

[20] Iria da Cunha, Juan-Manuel Torres-Moreno, and Gerardo Sierra. On the Development of the RST Spanish Treebank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 1–10, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[21] Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[22] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. AugGPT: Leveraging ChatGPT for Text Data Augmentation, 2023.

[23] Laurence Danlos, Diégo Antolinos-Basso, Chloé Braud, and Charlotte Roze. Vers le FDTB : French Discourse Tree Bank (Towards the FDTB : French Discourse Tree Bank) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 471–478, Grenoble, France, June 2012. ATALA/AFCP.

[24] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From Game Design Elements to Gamefulness: Defining Gamification. volume 11, pages 9–15, 09 2011. doi: 10.1145/2181037.2181040.

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423.

[26] William B. Dolan and Chris Brockett. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[27] Yulia Yu. Dyulicheva and Anastasia O. Glazieva. Game based learning with artificial intelligence and immersive technologies: an overview. In *4th Workshop for Young Scientists in Computer Science Software Engineering*, pages 146–159, December 2021.

[28] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What's In My Big Data?, 2023.

[29] Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. Ranking Generated Summaries by Correctness: An Interesting but Challenging Application for Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1213.

[30] Christiane Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, Cambridge, MA, 1998. doi: 10.7551/mitpress/7287.001.0001.

[31] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552.

[32] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. Shortcut learning

in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 11 2020. doi: 10.1038/s42256-020-00257-z.

[33] Mor Geva, Yoav Goldberg, and Jonathan Berant. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1107.

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

[35] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017.

[36] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740.

[37] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020.

[38] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep Learning Scaling is Predictable, Empirically, 2017.

[39] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735.

[40] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1243.

[41] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392. Association for the Advancement of Artificial Intelligence, May 2021. doi: 10.1609/aaai.v35i7.16792.

[42] Kiyoshi Izumi and Hiroki Sakaji. Economic Causal-Chain Search using Text Mining Technology. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 61–65, Macao, China, August 2019.

[43] Yangfeng Ji and Jacob Eisenstein. One Vector is Not Enough: Entity-Augmented Distributed Semantics for Discourse Relations. *Transactions of the Association for Computational Linguistics*, 3:329–344, 2015. doi: 10.1162/tacl_a_00142.

[44] Feng Jiang, Sheng Xu, Xiaomin Chu, Peifeng Li, Qiaoming Zhu, and Guodong Zhou. MCDTB: A Macro-level Chinese Discourse TreeBank. In *Proceedings of the 27th International Conference on Computational Linguis-*

tics, pages 3493–3504, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[45] Feng Jiang, Yaxin Fan, Xiaomin Chu, Peifeng Li, and Qiaoming Zhu. Not Just Classification: Recognizing Implicit Discourse Relation on Joint Modeling of Classification and Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2418–2431, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.187.

[46] Yuxin Jiang, Linhan Zhang, and Wei Wang. Global and Local Hierarchy-aware Contrastive Framework for Implicit Discourse Relation Recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8048–8064, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.510.

[47] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[48] Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. When Choosing Plausible Alternatives, Clever Hans can be Clever. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 33–42, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6004.

[49] Daisuke Kawahara and Sadao Kurohashi. Case Frame Compilation from the Web using High-Performance Computing. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA).

[50] Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. Rapid Development of a Corpus

with Discourse Annotations using Two-stage Crowdsourcing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 269–278, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[51] Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. 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, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/E14-1007.

[52] Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. Implicit Discourse Relation Classification: We Need to Talk about Evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.480.

[53] Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Improving Crowdsourcing-Based Annotation of Japanese Discourse Relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA).

[54] Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. Adapting BERT to Implicit Discourse Relation Classification with a Focus on Discourse Connectives. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1152–1158, Marseille, France, May 2020. European Language Resources Association.

[55] Yudai Kishimoto, Murawaki Yugo, Daisuke Kawahara, and Sadao Kurohashi. Japanese Discourse Relation Analysis: Task Definition, Connective Detection, and Corpus Annotation. *Journal of Natural Language Processing*, 27(4):889–931, 2020. doi: 10.5715/jnlp.27.889. (in Japanese).

[56] Hirokazu Kiyomaru. Studies on Fundamental Problems in Event-Level Language Analysis. Technical report, Japan, 2022.

[57] Hirokazu Kiyomaru, Nobuhiro Ueda, Takashi Kodama, Yu Tanaka, Ribeka Tanaka, Daisuke Kawahara, and Sadao Kurohashi. CausalityGraph: A System to Organize Causes, Results, and Solutions of Events based on Structural Language Analysis. In *Proceedings of 26th Annual Conference of Association for Natural Language Processing*, pages 1125—-1128, Online, March 2020. Association for Natural Language Processing. (in Japanese).

[58] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese General Language Understanding Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, June 2022. European Language Resources Association (ELRA).

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

[60] Wendy G. Lehnert. *The Process of Question Answering*. Routledge, London, first edition, 1978. doi: 10.4324/9781003316817.

[61] Douglas B. Lenat. CYC: A Large-Scale Investment in Knowledge Infrastructure. *Commun. ACM*, 38(11):33–38, November 1995. ISSN 0001-0782. doi: 10.1145/219717.219745.

[62] Hector J. Levesque. The Winograd Schema Challenge. In *the 2011 AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 63–68. Association for the Advancement of Artificial Intelligence, 2011.

[63] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages

1823–1840, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165.

[64] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.508.

[65] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10.

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

[67] Wanqiu Long and Bonnie Webber. Facilitating Contrastive Learning of Discourse Relational Senses by Exploiting the Hierarchy of Sense Relations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10704–10716, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.734.

[68] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13480–13488. Association for the Advancement of Artificial Intelligence, 2021. doi: 10.1609/aaai.v35i15.17590.

[69] Bill MacCartney. Natural Language Inference. Technical report, USA, 2009.

[70] Bill MacCartney. Understanding Natural Language Understanding. In *ACM SIGAI Bay Area Chapter Inaugural Meeting*, 2014.

[71] WILLIAM C. MANN and SANDRA A. THOMPSON. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281, 1988. doi: doi:10.1515/text.1.1988.8.3.243.

[72] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning Language Models as Training Data Generators for Augmentation-Enhanced Few-Shot Learning, 2023.

[73] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

[74] William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. Provable limitations of acquiring meaning from ungrounded form: What will future language models understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060, 2021. doi: 10.1162/tacl_a_00412.

[75] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759.

[76] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological Analysis for Unsegmented Languages using Recurrent Neural Network Language Model. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2297, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1276.

[77] Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations*, May 2021.

[78] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098.

[79] Yusuke Muzitani, Daisuke Kawahara, and Sadao Kurohashi. Construction of a Word Database based on Educational Level Specific Words using Crowdsourcing Questionnaires. In *Proceedings of the 25th Annual Meeting of the Association for Natural Language Processing*, pages 1503–1506, Nagoya, Aichi, March 2019. Association for Natural Language Processing.

[80] Makoto Nagao. Machine Translation Through Language Understanding. In *Proceedings of Machine Translation Summit VI: Plenaries*, pages 41–49, San Diego, California, October 29 – November 1 1997.

[81] National Institute for Educational Policy Research. Survey on a Particular Subject (Japanese) Results, 2008. URL `https://www.nier.go.jp/kaihatsu/tokutei/04002010000004000.pdf`. (in Japanese).

[82] Allen Nie, Erin Bennett, and Noah Goodman. DisSent: Learning Sentence Representations from Explicit Discourse Relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1442.

[83] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association*

for Computational Linguistics, pages 4885–4901, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.441.

[84] OpenAI. GPT-4 Technical Report, 2023.

[85] Naoki Otani, Hirokazu Kiyomaru, Daisuke Kawahara, and Sadao Kurohashi. Cross-lingual Knowledge Projection Using Machine Translation and Target-side Knowledge Base Completion. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1508–1520, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[86] Umangi Oza, Rashmi Prasad, Sudheer Kolachina, Dipti Misra Sharma, and Aravind Joshi. The Hindi Discourse Relation Bank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 158–161, Suntec, Singapore, August 2009. Association for Computational Linguistics.

[87] Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. Discourse Marker Augmented Network with Reinforcement Learning for Natural Language Inference. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 989–999, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1091.

[88] Aarthi Paramasivam and S. Jaya Nirmala. A survey on textual entailment based question answering. *Journal of King Saud University - Computer and Information Sciences*, 34(10, Part B):9644–9653, 2022. doi: https://doi.org/10.1016/j.jksuci.2021.11.017.

[89] Jason Phang, Thibault Févry, and Samuel R. Bowman. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks, 2018.

[90] Lucie Poláková, Jiří Mírovský, Anna Nedoluzhko, Pavlína Jínová, Šárka Zikánová, and Eva Hajičová. Introducing the Prague Discourse Treebank 1.0. In *Proceedings of the Sixth International Joint Conference on Natural*

Language Processing, pages 91–99, Nagoya, Japan, October 2013. Asian Federation of Natural Language Processing.

[91] Rashmi Prasad, Aravind Joshi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, and Bonnie Webber. The Penn Discourse TreeBank as a Resource for Natural Language Generation. In *Proceedings of the Corpus Linguistics Workshop on Using Corpora for Natural Language Generation*, pages 25–32, Birmingham, UK, July 2005.

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

[93] Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, and Bonnie Webber. Penn Discourse Treebank Version 2.0, 2008.

[94] Rashmi Prasad, Bonnie Webber, Alan Lee, and Aravind Joshi. Penn Discourse Treebank Version 3.0, 2019.

[95] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.467.

[96] Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. Training Question Answering Models From Synthetic Data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.468.

[97] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.14.

[98] Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1006–1017, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1093.

[99] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[100] Altaf Rahman and Vincent Ng. Resolving Complex Cases of Definite Pronouns: The Winograd Schema Challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[101] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264.

[102] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124.

[103] Ritsumeikan University Library. Detail of Reference Example, 2017. URL `https://crd.ndl.go.jp/reference/detail?page=ref_view&id=1000209543`. (in Japanese).

[104] Melissa Roemmele, Cosmin Bejan, and Andrew Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *the 2011 AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95. Association for the Advancement of Artificial Intelligence, 2011.

[105] Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A Constructive Prediction of the Generalization Error Across Scales. In *International Conference on Learning Representations*, 2020.

[106] Sebastian Ruder. Why You Should Do NLP Beyond English, 2020. URL `http://ruder.io/nlp-beyond-english`.

[107] Attapol Rutherford and Nianwen Xue. Improving the Inference of Implicit Discourse Relations via Classifying Explicit Discourse Connectives. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 799–808, Denver, Colorado, May–June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1081.

[108] Jun Saito, Tomohiro Sakaguchi, Tomohide Shibata, Daisuke Kawahara, and Sadao Kurohashi. Jutsugo Ko Kozo ni Motozuku Gengo Joho no Kihon Tan'i no Dezain to Kashika. In *Proceedings of the 24th Annual Meeting of the Association for Natural Language Processing*, pages 93–96, Okayama, Japan, March 2018. The Association for Natural Language Processing. URL `https://www.anlp.jp/proceedings/annual_meeting/2018/pdf_dir/E1-1.pdf`. (in Japanese).

[109] Jun Saito, Yugo Murawaki, and Sadao Kurohashi. Minimally Supervised Learning of Affective Events Using Discourse Relations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5758–5765, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1581.

[110] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740. Association for the Advancement of Artificial Intelligence, Apr. 2020. doi: 10.1609/aaai.v34i05.6399.

[111] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035. Association for the Advancement of Artificial Intelligence, Jul. 2019. doi: 10.1609/aaai.v33i01.33013027.

[112] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454.

[113] Timo Schick and Hinrich Schütze. Generating Datasets with Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.555.

[114] Ming Shen, Pratyay Banerjee, and Chitta Baral. Unsupervised Pronoun Resolution via Masked Noun-Phrase Prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the*

11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 932–941, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.117.

[115] Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. Nihon Go Winograd Schema Challenge no Kochiku to Bunseki. In *Proceedings of the Twenty-first Annual Meeting of the Association for Natural Language Processing*, Kyoto, Japan, March 2015. The Association for Natural Language Processing. URL `https://www.anlp.jp/proceedings/annual_meeting/2015/pdf_dir/E3-1.pdf`. (in Japanese).

[116] Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. Open Mind Common Sense: Knowledge Acquisition from the General Public. In *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-36124-4. doi: 10.1007/3-540-36124-3_77.

[117] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[118] John F. Sowa. Semantic Networks. *Encyclopedia of Artificial Intelligence*, 1992.

[119] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pages 4444–4451. Association for the Advancement of Artificial Intelligence, Feb. 2017. doi: 10.1609/aaai.v31i1.11164.

[120] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya

Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, ..., and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, 2023.

[121] Ieva Staliunaite, Philip John Gorinski, and Ignacio Iacobacci. Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13834–13842. Association for the Advancement of Artificial Intelligence, 2021. doi: 10.1609/aaai.v35i15.17630.

[122] Manfred Stede. The Potsdam Commentary Corpus. In *Proceedings of the Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[123] Manfred Stede and Arne Neumann. Potsdam Commentary Corpus 2.0: Annotation for Discourse Research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[124] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421.

[125] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[126] Alexandre Tamborrino, Nicola Pellicanò, Baptiste Pannier, Pascal Voitot, and Louise Naudin. Pre-training Is (Almost) All You Need: An Application to Commonsense Reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3878–3887, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.357.

[127] Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. From Discourse to Narrative: Knowledge Projection for Event Relation Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 732–742, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.60.

[128] Arseny Tolmachev, Daisuke Kawahara, and Sadao Kurohashi. Juman++: A Morphological Analysis Toolkit for Scriptio Continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 54–59, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2010.

[129] Alan M. Turing. Computing Machinery and Intelligence. *Mind*, 59(236): 433–460, 1950. doi: 10.1093/mind/lix.236.433.

[130] Sylke Vandercruysse, Mieke Vandewaetere, and Geraldine Clarebout. *Game-Based Learning: A Review on the Effectiveness of Educational Games*, volume 1, pages 628–647. 02 2012. doi: 10.4018/978-1-4666-0149-9.ch032.

[131] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[132] Nynke Vliet, Ildikó Berzlánovich, Gosse Bouma, Markus Egg, and Gisela

Redeker. *Building a Discourse-Annotated Dutch Text Corpus*, volume 3, pages 157–171. 01 2011.

[133] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2018.

[134] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, volume 32, pages 3266–3280. Curran Associates, Inc., 2019.

[135] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural Network Acceptability Judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290.

[136] Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. The Penn Discourse Treebank 3.0 Annotation Manual, 2019. URL `https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf`.

[137] Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1363.

[138] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[139] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101.

[140] Terry Winograd. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. Technical report, USA, 1971.

[141] Florian Wolf and Edward Gibson. Representing Discourse Coherence: A Corpus-Based Study. *Computational Linguistics*, 31(2):249–287, 2005. doi: 10.1162/0891201054223977.

[142] Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. A Label Dependence-Aware Sequence Generation Model for Multi-Level Implicit Discourse Relation Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11486–11494, Jun. 2022. doi: 10.1609/aaai.v36i10.21401.

[143] Wei Xiang and Bang Wang. A Survey of Implicit Discourse Relation Recognition. *ACM Comput. Surv.*, 55(12), mar 2023. doi: 10.1145/3574134.

[144] Wei Xiang, Zhenglin Wang, Lu Dai, and Bang Wang. ConnPrompt: Connective-cloze Prompt Learning for Implicit Discourse Relation Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 902–911, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[145] Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 725–731, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1079.

[146] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative Data Augmentation for Commonsense Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.90.

[147] Zhi-Xiu Ye, Qian Chen, Wen Wang, and Zhen-Hua Ling. Align, Mask and Select: A Simple Method for Incorporating Commonsense Knowledge into Language Representation Models, 2019. URL `http://arxiv.org/abs/1908.06725`.

[148] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 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, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1009.

[149] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472.

[150] Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences*, 10(21):7640, 2020. doi: 10.3390/app10217640.

[151] Deniz Zeyrek and Bonnie Webber. A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. In *Proceedings of the 6th Workshop on Asian Language Resources*, 2008.

[152] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. ASER: A Large-Scale Eventuality Knowledge Graph. In *Proceedings*

of The Web Conference 2020, page 201–211, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3366423.3380107.

[153] Hao Zhou, Man Lan, Yuanbin Wu, Yuefeng Chen, and Meirong Ma. Prompt-based Connective Prediction Method for Fine-grained Implicit Discourse Relation Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3848–3858, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.282.

[154] Yuping Zhou and Nianwen Xue. The Chinese Discourse Treebank: a Chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49:397–431, 06 2015. doi: 10.1007/s10579-014-9290-3.

List of Major Publications

[1] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. A Method for Building a Commonsense Inference Dataset based on Basic Events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2450–2460, 2020.

[2] Kazumasa Omura and Sadao Kurohashi. Improving Commonsense Contingent Reasoning by Pseudo-data and its Application to the Related Tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 812-823, 2022. (**Outstanding Paper Award**)

[3] Kazumasa Omura, Kei Kubo, Frederic Bergeron, and Sadao Kurohashi. Toward Game-Based Learning of Japanese Writing for Elementary School Students. In *Proceedings of the 31st International Conference on Computers in Education (ICCE)*, pages 655–660, 2023.

[4] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. Building a Commonsense Inference Dataset based on Basic Events and its Application. *Journal of Natural Language Processing*, 30(4), pages 1206-1239, 2023. (in Japanese, **Best Paper Award**)

[5] Kazumasa Omura, Fei Cheng, and Sadao Kurohashi. An Empirical Study of Synthetic Data Generation for Implicit Discourse Relation Recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024. (to appear)

List of Other Publications

[6] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. Building a Commonsense Inference Dataset based on Basic Events. In *Proceedings of 26th Annual Meeting of Association for Natural Language Processing*, pages 1539-1542, 2020. (in Japanese)

[7] Kazumasa Omura. A Method for Building a Commonsense Inference Dataset based on Basic Events. *Journal of Natural Language Processing*, 28(1), pages 287-291, 2021. (in Japanese)

[8] Kazumasa Omura, Kei Kubo, and Sadao Kurohashi. Word Connect Game: Gamification of Japanese Writing Education for Elementary School Students. In *Proceedings of 27th Annual Meeting of Association for Natural Language Processing*, pages 895-899, 2021. (in Japanese)

[9] Kazumasa Omura and Sadao Kurohashi. Improving Commonsense Contingent Reasoning by Pseudo-data and its Application to the Related Tasks. In *Proceedings of 28th Annual Meeting of Association for Natural Language Processing*, pages 2029-2034, 2022. (in Japanese)

[10] Kazumasa Omura, Hono Shirai, Shotaro Ishihara, and Norihiko Sawa. Automatic Generation of Training Data from Financial Report Briefing Articles for Automatic Extraction of Financial Factors from Quarterly Financial Reports. In *Proceedings of 28th Annual Meeting of Association for Natural Language Processing*, pages 1449-1454, 2022. (in Japanese)

[11] Kazumasa Omura, Hono Shirai, Shotaro Ishihara, and Norihiko Sawa. Automatic Extraction of Financial Factors from Quarterly Financial Reports Considering Their Polarity and Importance. In *Proceedings of 29th Annual Meeting of Association for Natural Language Processing*, pages 2703-2708, 2023. (in Japanese)

[12] Hirokazu Kiyomaru, Kazumasa Omura, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Diversity-aware Event Prediction based on a Conditional Variational Autoencoder with Reconstruction. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 113-122, 2019.

[13] Akiko Aizawa, Frederic Bergeron, Junjie Chen, Fei Cheng, Katsuhiko Hayashi, Kentaro Inui, Hiroyoshi Ito, Daisuke Kawahara, Masaru Kitsuregawa, Hirokazu Kiyomaru, Masaki Kobayashi, Takashi Kodama, Sadao Kurohashi, Qianying Liu, Masaki Matsubara, Yusuke Miyao, Atsuyuki Morishima,

Yugo Murawaki, Kazumasa Omura, Haiyue Song, Eiichiro Sumita, Shinji Suzuki, Ribeka Tanaka, Yu Tanaka, Masashi Toyoda, Nobuhiro Ueda, Honai Ueoka, Masao Utiyama, and Ying Zhong. A System for Worldwide COVID-19 Information Aggregation. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, 2020.

[14] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A Unified Japanese Analyzer Based on Foundation Models. In *Proceedings of the 253-th Special Interest Group of Natural Language Processing*, pages 1-14, 2022. (in Japanese, **Outstanding Research Award**)

[15] Takashi Kodama, Nobuhiro Ueda, Kazumasa Omura, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Japanese Morphological Analysis Using Text Generation Model. In *Proceedings of 29th Annual Meeting of Association for Natural Language Processing*, pages 339-344, 2023. (in Japanese)

[16] Nobuhiro Ueda, Kazumasa Omura, Takashi Kodama, Hirokazu Kiyomaru, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. KWJA: A Unified Japanese Analyzer Based on Foundation Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 538-548, 2023.