疑似問題による常識推論能力の改善と関連タスクへの効果

大村 和正　黒橋 禎夫
京都大学大学院情報学研究科
{omura,kuro}@nlp.ist.i.kyoto-u.ac.jp

概要

本研究では，常識推論の中でも基本的な出来事の間の蓋然的関係を推論する能力の改善に取り組む．既存のデータ構築手法をベースに，テキストから蓋然的関係を持つ基本的なイベント表現の組を自動抽出し，これをもとに疑似問題を大規模に自動生成することで，常識推論能力の改善を試みる．実験の結果，疑似問題により常識推論タスクの性能が向上すること，蓋然的関係知識を転移することで談話関係解析に効果があることを示した．

1　はじめに

人間は文章を読む時，記述された出来事から引き起こされる状況を推論することで内容の理解を深める．また，対話においては，相手の発話から次の展開を推論することで文脈に沿った自然な応答を返す．このように，人間は観測される出来事と蓋然的関係を持つ出来事の推論を日常的に行っており，これは言語理解において重要な能力のひとつである．

蓋然的関係 (Contingency) は，ある事象に対してある程度次に起こりうる事象との間に成立する関係である．Penn Discourse Tree Bank[1] においては主要な談話関係のひとつとして扱われており，因果関係や前提条件などの談話関係を下位分類に含む．

近年，計算機による言語理解の実現に向けて，この蓋然的推論に焦点を当てた言語資源が盛んに構築されている [2, 3, 4, 5, 6]．これらの言語資源は基本的な出来事に対象を絞り，ある種の常識推論能力を問う．深層学習モデルの発展に伴い，自然言語推論をはじめとする基礎的な言語理解力は飛躍的に向上しているものの，計算機は人間と比べてこの常識推論能力に乏しいことが実験的に示されている．

本研究では，京都大学常識推論データセット (KUCI)[1)] を利用し，常識推論能力の改善に取り組む．KUCI は日本語の QA データセットのひとつであり，図 1 のような，基本的な出来事の間の蓋然的関係を問う多肢選択式問題約 10 万問から成る [7]．ウェブテキストから原因と結果の関係を持つ基本的なイベント表現の組を抽出し，クラウドソーシングで確認した後，これをベースに問題を生成するという半自動的な構築手法が特徴である．

お腹が空いたので
1. 学校休みます
2. 激しい運動は控えます
3. ファミレスで食事する
4. 私は家を出ます

図 1　KUCI に含まれる常識推論問題の例．太字は正解選択肢である．

このタスクにおいても，計算機と人間の間に正答率の開きがあることが示されている．また，定性評価の結果，計算機はごく基本的なイベント間の蓋然的関係についても誤答している例が散見された．この問題は素朴には，訓練データを拡張し，カバレッジを向上させることで軽減できると考えられる．しかし，データ構築時にクラウドソーシングを利用しているため，同様の手法でデータを倍々に拡張していくことはコスト面から現実的でない．

我々は，データ拡張のボトルネックとなっているクラウドソーシングを省略し，正解が保証されていない疑似問題を組み合わせることで改善を試みる．抽出元のウェブテキストは容易に拡張可能であり，クラウドソーシング以外は自動処理であるため，大規模な疑似問題を生成することができる．これにより，訓練データのカバレッジ不足を補う．

蓋然的推論が言語理解において重要であるならば，蓋然的関係知識を転移することでその他のタスクにおいても改善が期待できる．主要な英語のデータセットは転移可能性が検証されてきた [8, 9, 10, 11] 一方で，この常識推論データセットについては探索の余地がある．我々は，転移学習による関連タスクへの効果を定量的に評価し，蓋然的関係知識の汎用性を検証する．

1）https://nlp.ist.i.kyoto-u.ac.jp/?KUCI

2 関連研究

大規模な事前学習の枠組みにより，常識推論を含む様々な自然言語処理タスクで先例のない性能が達成されている [12]．このような汎用的な言語理解力の改善とは別に，常識推論能力の改善に向けたアプローチもこれまでに数多くとられてきた．

ひとつは自動生成した訓練データを利用するもので，我々のアプローチもこれに該当する．例えば，Ye らは，Wikipedia と ConceptNet[13] から自動生成した 1,600 万問規模の多肢選択式穴埋め問題に対して追加の事前学習を行った [14]．人手で構築された言語資源 (ConceptNet) を必要とするものの，エンティティレベルの常識推論タスク (CommonsenseQA[15] および Winograd Schema Challenge[16]) において性能の向上が確認されている．Staliunaite らは，ウェブテキストから談話標識を手がかりに因果関係を持つ節の組を抽出し，負例を言語モデルから生成することで，COPA[2] に対するデータ拡張を行った [17]．常識的な因果関係の推論能力の改善に焦点を当てており，関連タスクへの影響は検証していない．

その他のアプローチとして，常識推論に向けた既存の言語資源から知識を転移するものがある．例えば，常識推論に向けたベンチマークの 1 つである Social IQA[9] および WinoGrande[10] を中間タスクとして解くことで，Winograd Schema Challenge および COPA に対する性能の向上が報告されている [9, 10]．Pruksachatkun らは，CosmosQA[18] や HellaSwag[19] といった複雑な常識推論を必要とするデータセットが転移元のタスクとして効果的であることを実験的に示した [11]．また，類似のアプローチとして，複数の常識推論データセットでマルチタスク学習 [20] を行う手法 [21]，常識知識ベースを外部知識として組み込む手法 [22] などが提案されている．

3 アプローチ

3.1 疑似問題の生成手法

我々のアプローチは，京都大学常識推論データセット (KUCI) の構築手法をベースにデータ拡張のボトルネックとなっているクラウドソーシングを省略することで，大規模な疑似問題を自動生成するというものである．KUCI に含まれる常識推論問題の生成手法は，大まかに以下の 4 ステップから成る (図 2)．

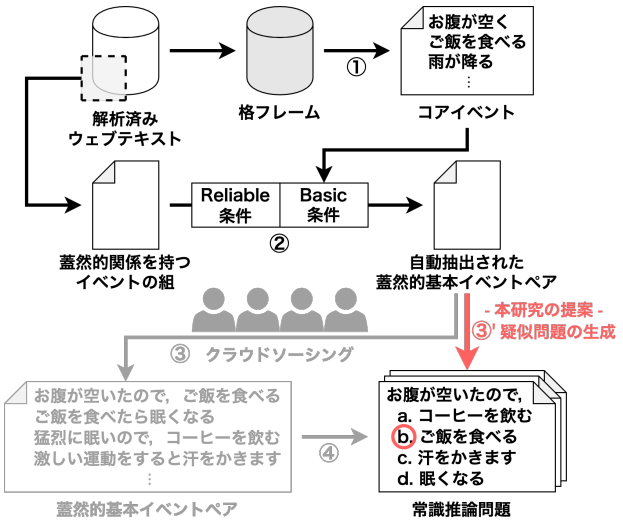

図 2 KUCI に含まれる常識推論問題と疑似問題の生成手法の概要図．

1. 格フレーム [23] から高頻度な述語項構造 (**コアイベント**と呼ぶ) を獲得する
2. 蓋然的関係を表す談話標識で接続され，節間に係り受けの曖昧性がなく，前件・後件ともにコアイベントを含む節の組 (**蓋然的基本イベントペア**と呼ぶ) をテキストから抽出する
3. 抽出されたイベントペアが蓋然的関係を持つか否かをクラウドソーシングで確認する
4. 蓋然的関係を持つと判断されたイベントペアをベースに，これと中程度に類似するイベントペアの後件から誤り選択肢を無作為に抽出することで問題を生成する

この手法において，ステップ 3 を省略することで問題の自動生成が可能となる (図 2 ③')．コアイベントを獲得する際の頻度に関する閾値や誤り選択肢の抽出条件など，手法のパラメータは KUCI の構築時と同じ値に設定する．

3.2 蓋然的基本イベントペアの自動抽出

3.1 節の手法に従って蓋然的基本イベントペアの自動抽出を行った．抽出元のテキストは，ウェブから収集した日本語 33 億文から成るコーパスを利用した．このコーパスと KUCI の構築時に利用したウェブコーパスとの間に文の重複はない．この結果，83.2 万組の蓋然的基本イベントペアが自動抽出された．Omura らの報告では，クラウドソーシングによって約 1/3 のイベントペアが除かれているため [7]，約 50 万組のイベントペアは有効であることが期待される．

3.3 リークへの対処

大規模テキストから訓練データを生成する際の懸念として Data Contamination の問題がある [24]．テキスト中に評価データと同一もしくは酷似した文が含まれているために意図せず教師信号を学習し，性能が過大評価されてしまうというものである．

本研究では，評価データに含まれる問題の生成元となったイベントペア (**ベース**と呼ぶ) と酷似するものをヒューリスティックにもとづき除去することで対処する．具体的には，単語の並びおよびコアイベントの組にもとづくフィルタリングを適用する．

単語の並びにもとづくフィルタリング
ベースと重複する単語の並びの長さが，ベースの単語数の 80%を超えるものを除く．

コアイベントの組にもとづくフィルタリング
ベースに含まれるコアイベントの組と同じコアイベントの組を含むイベントペアを除く．

上記のフィルタリングを適用した結果，77.4 万組の蓋然的基本イベントペアが獲得された．これらのイベントペアから問題を生成した結果，77.2 万問の疑似問題が生成された．無作為に選択した 50 問を著者が評価した結果，36 問が解答可能であった．

4 計算機による解答実験

疑似問題による効果を検証するため，深層学習モデルによる KUCI および関連タスクの解答実験を行った．本研究では，関連タスクとして談話関係解析および日本語 Winograd Schema Challenge (JWSC)[25] を対象とした．

4.1 モデル

本研究では，BERT モデル [26] および XLM-RoBERTa(XLM-R) モデル [27] の性能を検証した．BERT は，日本語 Wikipedia 全文で事前学習した NICT BERT 日本語 Pre-trained モデル (BPE あり)[2] を利用した．XLM-R は，Wikipedia および CC-100[28] から成る大規模多言語コーパスで事前学習した XLM-RoBERTa$_{\text{LARGE}}$ モデル[3] を利用した．

4.2 実験設定

常識推論　前述のとおり，KUCI を用いて常識推論能力を評価した．タスクは 4 択問題で，訓練用，開発用，テスト用にそれぞれ 83,127 問，10,228 問，10,291 問を含む．各選択肢のスコアは，文脈と選択肢の組を特殊な記号で区切って入力し，先頭トークンのベクトル表現をスカラーに線形変換することで算出する．訓練時は，softmax 関数で正規化した各選択肢のスコアと正解選択肢を 1 とする one-hot vector の間の交差エントロピー誤差を最小化するように学習する．なお，疑似問題を加える時は，常識推論問題の交差エントロピー誤差 L_{CI} と疑似問題の交差エントロピー誤差 L_{Pseudo} の重み付き和

$$L = L_{CI} + \lambda L_{Pseudo}$$

を最小化する．

談話関係解析　データセットは京都大学ウェブ文書リードコーパス (KWDLC)[29, 30] を利用した．KWDLC は様々な文書の冒頭 3 文をウェブから収集することで構築されており，その規模は 6,445 文書から成る．これらの文書には，クラウドソーシングを用いて節間に談話関係がラベル付けされている．さらに，このうち 500 文書には専門家によるラベルも付与されている．本研究では，専門家ラベルのない節ペア 37,491 組を訓練データに利用し，専門家ラベルが付与された節ペア 2,320 組の分類精度を 5 分割交差検証で評価した．タスクは「談話関係なし」を含む 7 値分類であり，Devlin らが提案した文ペア分類の設定 [26] で fine-tuning した．

JWSC　WSC は文中の照応詞が指示する先行詞を 2 つの候補から選択するタスクである [16]．JWSC は訓練データ 1,322 文とテストデータ 564 文から成り，開発データは用意されていないため 5 分割交差検証を行った．先行詞に含まれるトークンのベクトル表現の平均をその先行詞のベクトル表現とみなし，正解の先行詞と照応詞のベクトル表現の間のコサイン類似度を 1 に近づけるように学習した．

表 1　KUCI の正解率．異なる 3 つのシード値で fine-tuning した結果の平均と標準偏差を記載している．

モデル	設定	正解率
BERT	KUCI	79.4 ± 0.1
	KUCI + 疑似問題 (λ = 0.1)	83.8 ± 0.3
	KUCI + 疑似問題 (λ = 0.5)	**84.9** ± **0.1**
	KUCI + 疑似問題 (λ = 1.0)	84.5 ± 0.3
XLM-R	KUCI	85.6 ± 0.4
	KUCI + 疑似問題 (λ = 0.1)	88.3 ± 0.2
	KUCI + 疑似問題 (λ = 0.5)	**88.5** ± **0.1**
	KUCI + 疑似問題 (λ = 1.0)	**88.5** ± **0.1**
クラウドワーカー [7]		88.9

2) https://alaginrc.nict.go.jp/nict-bert/index.html
3) https://huggingface.co/xlm-roberta-large

やはりメダルを獲れば	霧が晴れると、
a. 応援に熱が入る	**a. 景色が素晴らしい**
b. 話題になる	b. 川の音がすごい
c. 代替出場が決定する	c. 雪遊びも楽しそうだ
d. 彼の大会になる	d. 写真写りがいまいちだ

図 3　疑似問題により BERT モデルが正答するようになった問題例．太字は正解選択肢であり，赤字は以前の誤答選択肢である．

表 2　KWDLC に対する談話関係解析の結果．評価指標は「談話関係なし」を除いた micro-average で算出した．5 分割交差検証を異なる 3 つのシード値で行い，その結果の平均と標準偏差を記載している．また，矢印は多段階の fine-tuning を表す．

モデル	設定	F 値
BERT	KWDLC	44.3 ± 0.4
	KUCI → KWDLC	45.5 ± 0.5
	KUCI + 疑似問題$_{\lambda=0.5}$ → KWDLC	**47.7** ± **0.8**
XLM-R	KWDLC	51.4 ± 0.9
	KUCI → KWDLC	50.8 ± 0.6
	KUCI + 疑似問題$_{\lambda=0.5}$ → KWDLC	**51.9** ± **1.1**
クラウドワーカー [31]		51.5

表 3　JWSC の正解率．5 分割交差検証を異なる 3 つのシード値で行い，その結果の平均と標準偏差を記載している．† は一部訓練実行時に学習が進まなかったこと (degenerate run[8, 11]) を表す．

モデル	設定	正解率
BERT	JWSC	**72.0** ± **1.7**
	KUCI → JWSC	71.4 ± 1.5
	KUCI + 疑似問題$_{\lambda=0.5}$ → JWSC	66.5 ± 2.0
XLM-R	JWSC	75.6 ± 10.7†
	KUCI → JWSC	75.6 ± 2.4
	KUCI + 疑似問題$_{\lambda=0.5}$ → JWSC	**77.5** ± **1.7**

4.3　実験結果

常識推論タスクの実験結果を表 1 に示す．疑似問題を加えることで，BERT モデルと XLM-R モデルともに正解率が 3〜4%向上している．また，図 4 にモデルの学習曲線を示す．疑似問題は KUCI の訓練データと比べて低品質であるものの，性能の向上に寄与することが分かる．

疑似問題で訓練したモデルが正答するようになった問題例を図 3 に示す．これらの基本的な蓋然的関係を問う問題に正答するようになったほか，全ての選択肢に低いスコアを付けて消去法的に誤り選択肢を選んでいたような問題に対する改善なども見られた．疑似問題によって KUCI の訓練データのカバレッジ不足が補われていると考えられる．

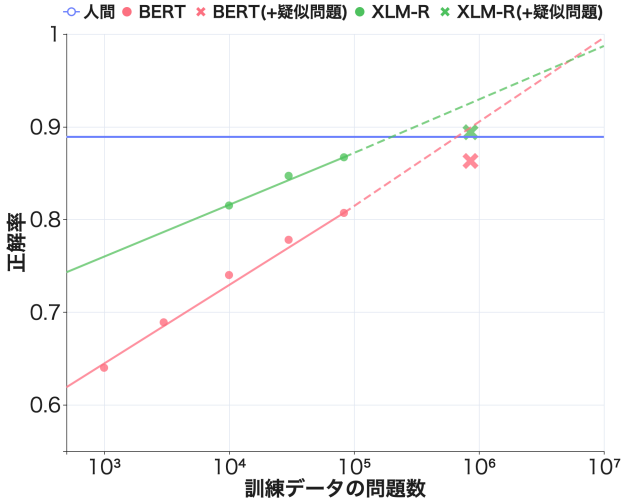

図 4　KUCI の開発データに対する BERT モデルと XLM-R モデルの学習曲線．訓練データ不足のために学習が進まなかった結果については省略している．

関連タスクについて，まず，表 2 に談話関係解析の実験結果を示す．KUCI および疑似問題を中間タスクとして解くことで，談話関係解析に効果があることが分かる．問題の生成元となっている蓋然的基本イベントペアは，KNP[32] によって「原因・理由」または「条件」の談話関係を持つと自動解析されたものであるため，これらの談話関係の知識が上手く転移されたと考えられる．

次に，JWSC に対する解答実験の結果を表 3 に示す．BERT モデルは，疑似問題を中間タスクとして解くことによる性能の悪化が見られる．照応先が文外にあるために理解が困難な問題が生成されることを防ぐ目的で，指示詞を含むイベントペアを除外しているため，訓練の途中で指示詞に関する知識が忘却されてしまうことが原因だと考えられる．一方で，XLM-R モデルにおいては，疑似問題による性能の悪化は見られなかったが，一部訓練時に学習が進まない現象 (degenerate run[8, 11]) が見られた．データが小規模である場合によく見られる現象であり，中間タスクによってこの問題を軽減できることが報告されている [8, 11] が，本研究でも同様の結果が確認された．

5　おわりに

本研究では，大規模に自動生成した疑似問題による常識推論能力の改善と，関連タスクへの影響の検証に取り組んだ．今後は，現状の計算機モデルが誤答する問題の定性的分析を進め，より深い言語理解を問うデータの構築を検討する．

謝辞

本研究は，(公財) 日本漢字能力検定協会の支援を受けた．

参考文献

[1] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In **Proc. LREC2008**.

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

[3] 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 **Proc. NAACL2016**.

[4] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In **Proc. EMNLP2018**.

[5] 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 **Proc. AAAI2019**.

[6] 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 **Proc. AAAI2021**.

[7] Kazumasa Omura, Daisuke Kawahara, and Sadao Kurohashi. A Method for Building a Commonsense Inference Dataset based on Basic Events. In **Proc. EMNLP2020**.

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

[9] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense Reasoning about Social Interactions. In **Proc. EMNLP2019**.

[10] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. In **Proc. AAAI2020**.

[11] 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 **Proc. ACL2020**.

[12] 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 **Proc. NeurIPS2019**.

[13] Robyn Speer, Joshua Chin, and Catherine Havasi. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In **Proc. AAAI2017**.

[14] 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. **CoRR**, Vol. abs/1908.06725, , 2019.

[15] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In **Proc. NAACL2019**.

[16] Hector J. Levesque. The Winograd Schema Challenge. In **Proc. AAAI2011 Spring Symposium**.

[17] Ieva Staliunaite, Philip John Gorinski, and Ignacio Iacobacci. Improving Commonsense Causal Reasoning by Adversarial Training and Data Augmentation. In **Proc. AAAI2021**.

[18] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In **Proc. EMNLP2019**.

[19] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In **Proc. ACL2019**.

[20] Rich Caruana. Learning Many Related Tasks at the Same Time with Backpropagation. In **Proc. NeurIPS1995**.

[21] Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. In **Proc. AAAI2021**.

[22] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. KagNet: Knowledge-Aware Graph Networks for Commonsense Reasoning. In **Proc. EMNLP2019**.

[23] Daisuke Kawahara, Daniel Peterson, Octavian Popescu, and Martha Palmer. Inducing Example-based Semantic Frames from a Massive Amount of Verb Uses. In **Proc. EACL2014**.

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

[25] 柴田知秀, 小浜翔太郎, 黒橋禎夫. 日本語 Winograd Schema Challenge の構築と分析. 言語処理学会第 21 回年次大会論文集 (NLP 2015).

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **Proc. NAACL2019**.

[27] 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 **Proc. ACL2020**.

[28] 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 **Proc. LREC2020**.

[29] 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 **Proc. COLING2014**.

[30] Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi. Improving Crowdsourcing-Based Annotation of Japanese Discourse Relations. In **Proc. LREC2018**.

[31] 岸本裕大, 村脇有吾, 河原大輔, 黒橋禎夫. 日本語談話関係解析：タスク設計・談話標識の自動認識・コーパスアノテーション. 自然言語処理, Vol. 27, No. 4, 2020.

[32] Sadao Kurohashi and Makoto Nagao. A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. **Computational Linguistics**, Vol. 20, No. 4, pp. 507–534, 1994.

表 7　アンサンブルモデルの談話関係ごとの True Positive および Positive の数．異なる 3 つのシード値で fine-tuning したモデルを Seed Averaging によってアンサンブルした結果を記載している．

モデル	設定	原因・理由	目的	条件	根拠	対比	逆接	適合率	再現率	F 値
BERT (ensemble)	KWDLC	87 / 155	18 / 33	28 / 38	0 / 8	3 / 16	47 / 76	56.1	40.4	47.0
	KUCI → KWDLC	80 / 140	16 / 31	32 / 44	1 / 4	1 / 10	53 / 79	59.4	40.4	48.1
	KUCI + 疑似問題$_{\lambda=0.5}$ → KWDLC	90 / 145	16 / 32	31 / 44	2 / 10	2 / 11	50 / 69	**61.4**	**42.2**	**50.0**
XLM-R (ensemble)	KWDLC	96 / 166	18 / 36	34 / 49	3 / 5	0 / 18	62 / 87	**59.0**	47.0	52.3
	KUCI → KWDLC	103 / 190	21 / 35	38 / 53	1 / 4	0 / 24	58 / 84	56.7	**48.8**	52.4
	KUCI + 疑似問題$_{\lambda=0.5}$ → KWDLC	104 / 185	19 / 36	36 / 51	0 / 1	1 / 24	61 / 89	57.3	**48.8**	**52.7**
True Positive と True Negative の合計		242	36	54	15	6	100	—		

A　実験の詳細

A.1　ハイパーパラメータ

fine-tuning に関するハイパーパラメータの詳細は以下のとおりである．

表 4　ハイパーパラメータの詳細 (KUCI).

パラメータ名	パラメータ値
バッチサイズ	32
エポック数	3
学習率	2e-5 (BERT) 5e-6 (XLM-R)
最大トークン長	128
Optimizer	AdamW
Adam's betas params	(0.9, 0.999)
Adam's epsilon	1e-6
重み減衰	1e-2
Scheduler	Linear decay with linear warmup
Warmup proportion	0.03
シード値	{0, 1, 2}

表 5　ハイパーパラメータの詳細 (KWDLC,JWSC).

パラメータ名	パラメータ値
バッチサイズ	32
エポック数	20
Patience for early stopping	5
学習率	2e-5 (BERT) 5e-6 (XLM-R)
最大トークン長	128
Optimizer	AdamW
Adam's betas params	(0.9, 0.999)
Adam's epsilon	1e-6
重み減衰	1e-2
Scheduler	Linear decay with linear warmup
Warmup proportion	0.03
シード値	{0, 1, 2}

XLM-R は学習率を低く設定することで学習が安定したため，BERT の学習率より低く設定している．

表 6　KUCI の開発データに対する BERT モデルの正答数・誤答数の混同行列．異なる 3 つのシード値で fine-tuning したモデルを Seed Averaging によってアンサンブルした結果を記載している．

		KUCI	
		正答	誤答
KUCI + 疑似問題 ($\lambda = 0.5$)	正答	7,902	979
	誤答	418	929

B　疑似問題の効果の詳細

B.1　常識推論問題の正答数の変化

KUCI の開発データに対する BERT モデルの正答数・誤答数の混合行列を表 6 に示す．

B.2　談話関係解析結果の詳細

談話関係ごとの分類結果を表 7 に示す．

C　フィルタリングの適用例

3.3 節のフィルタリングの適用例を述べる．例えば，図 1 の問題のベースは「お腹が空いたので → ファミレスで食事する」であり，「お腹が空く → ファミレスで食事する」というコアイベントの組を含む．これに対し，イベントペア「お腹が空いたので → 友達とファミレスで食事する」は{お腹, が, 空いた, ので, ファミレス, で, 食事, する}という 8 単語の並びが重複し，ベースの単語数 (8) の 80%を超える．また，同じコアイベントの組を含むため，両方のフィルタリングによって除かれる．