2020 年 3 月に言語処理学会年次大会で発表した本論文には、設定等の誤りによる実験結果の誤りがありました．
本論文中の数値は以下のものが正しい数値です．

- 条件 **Similarity** の類似度範囲: 0.5 より小さい → 0.4 から 0.6
- BERT-LARGE の解答精度: 39% → 79.5%
- 人間の解答精度: 92% → 94%

なお、論文の主張等に変更はございません．心よりお詫び申し上げます．

基本イベントに基づく常識推論データセットの構築

大村 和正　　河原 大輔　　黒橋 禎夫
京都大学 大学院情報学研究科

{omura, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

深層学習の進展にともない，計算機による言語理解力を鍛える／評価する問題設定を考え，そのデータを構築する研究が盛んである．

従来から計算機による言語理解には，言語に関する知識（語の意味，構文など）と，言語を超えた我々の世界・社会に関する知識（その基本的なものがいわゆる常識）の両方が必要であると議論されてきた．

このうち，前者の言語に関する知識の問題は，大規模テキストからの汎用言語モデル BERT [1] などによって相当程度解決した．文脈に応じた語の意味をベクトル表現できるようになり，それをもとにした fine-tuning によって，構文解析，パラフレーズ認識，深い推論を必要としない質問応答などについては人間に匹敵する精度が達成されている．

一方，後者の常識の獲得についてはまだ課題が多い．最大の問題は，常識というもののある種の汎用性を担保する必要があるという点にある．常識の汎用性とは，知識として基本的なものであると言い換えることもできる．さらに，もう一つの問題として，問題をつくる際のバイアスをできるだけ排除しなければならないという問題がある [2]．

本研究では，これらの問題を解決するためにテキストを利用する．蓋然性をもつ（＝よく起こり得る）基本的なイベントペア表現をテキストから抽出し，これをクラウドソーシングで確認することで問題を作るという方法を提案する．基本的なイベント表現（**基本イベント**と呼ぶ）は，テキストから抽出した述語項構造をクラスタリングし，その中の高頻度なものを核とする表現と定義する．そして，談話関係解析を用いることにより，蓋然性をもつ基本イベントペアを抽出する．これを**蓋然的基本イベントペア**と呼ぶ．

たとえば，抽出される蓋然的基本イベントペアは次のようなものである．

(1)　a.　お腹が空いたので，ご飯を食べる
　　b.　ご飯を食べたら，眠くなる
　　c.　眠いので，コーヒーを飲む
　　d.　激しい運動をすると，汗をかく

お腹が空いたので
1. コーヒーを飲む
2. ご飯を食べる
3. 汗をかく
4. 眠くなる

表 1:　常識推論問題の作問例 (太字は正解選択肢)

これらをベースとして，異なるペアの後件を誤り選択肢とすることで表 1 のような**常識推論問題**をつくることができる．

このようにテキストからの抽出をベースとするため，スケーラブルであり，ドメインが限定されることもない．クラウドソーシングについても，作文をするのではなくて確認・フィルタリングを行うだけなので，ここからバイアスが生まれることもない．

本研究では，提案手法に基づいてテキストから常識推論データセットを構築した．実験の結果，バイアスが小さなデータセットになっていることを確認した．

2　関連研究

これまでに構築された常識に関するデータセットとして，大規模なものでは SWAG [3] と CommonsenseQA [4] が挙げられる．

SWAG は，与えられた文脈に続く最も適切な動詞句を問う多肢選択式問題 11 万問からなる常識推論データセットである．常識としての汎用性を担保するため，動画キャプションから問題を作成しており，問題のドメインは物理世界に限定される．各問題は，動画キャプションから連続する 2 文を抽出し，その 1 文目から 2 文目の主語までを文脈，残りの部分を正解選択肢としている．誤り選択肢の候補は言語モデルから生成す

るため，言語モデルの生成バイアスに対処しなければならない．

CommonsenseQA は，問題文の答えとして最も適切な単語を問う多肢選択式問題 1.2 万問からなる常識問題データセットである．常識データベースである ConceptNet から部分グラフを抽出し，それを基にクラウドソーシングで問題を作成している．この手法は既存のリソースを利用するため，スケーラビリティに欠ける．また，問題文の作成はクラウドソーシングで行うため，annotation artifacts [2] の問題がある．

3　常識推論データセットの構築

常識推論問題は，表 1 に示すように，文脈と 4 つの選択肢からなり，文脈の後に続く文として最も適切なものを選択肢から選ぶ問題である．

これらの問題は，常識としての汎用性を担保するため，基本イベントに関するものに限定する．また，スケーラビリティの確保とバイアスの低減のために，コーパスからの自動抽出とクラウドソーシングによる確認を組み合わせる．これらの点を考慮し，次の手順で常識推論問題を生成する (図 1)．

1. 高頻度な述語項構造から基本イベントを獲得する．
2. コーパスに談話関係解析を適用し，その結果，蓋然的関係をもつと認識され，かつ基本イベントを核とするイベントペアを抽出する．
3. 抽出したイベントペアが蓋然的関係をもつかどうかをクラウドソーシングで確認し，蓋然的基本イベントペアを得る．
4. 蓋然的基本イベントペアから正解選択肢を作り，それ以外のイベントペアから良質な誤り選択肢を選択することによって，常識推論問題を生成する．

以下では各ステップの詳細について述べる．

3.1　基本イベントの獲得

本研究における基本イベントは，テキストから抽出した述語項構造を用法ごとにクラスタリングし，その中の高頻度なものを核とする表現とする．述語項構造をクラスタリングした既存のリソースとして，格フレーム[1] があり，本研究ではこれを利用する．

格フレームにおいては，各述語が用法ごとに複数の格フレームをもつ．各格フレームは複数の格をもち，各格は格要素となりうる名詞群からなる．日本語格フレームの例を表 2 に示す．

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

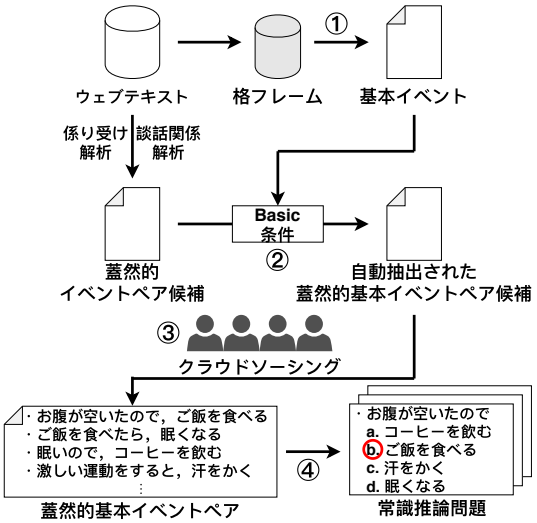

図 1:　提案手法の概要．

格フレーム	格	項
壊す (1)	ガ	私:83，人:65，...
	ヲ	お腹:25,643，体:17,242...
	デ	ストレス:297，食べ:174，...
壊す (2)	ガ	方:42，日本:42，...
	ヲ	雰囲気:8,140，イメージ:3,774...
	ノ	場:873，部屋:851，...
...		

表 2:　述語「壊す」の格フレーム．数値は頻度を表す．

本研究では，格フレームから高頻度な述語項構造を抽出し，それらを基に基本イベントを獲得する．まず，格フレームデータから，述語として，能動態のみを対象に頻度上位 5,000 件を取得する．取得した各述語に対して，格フレーム，格，格要素をそれぞれ頻度順に見たときに，頻度の累積和がその項目全体の 75%，50%，50%を越えるまで取得する．例えば，格フレームはその述語頻度の 75%をカバーするまで取得する．

この結果，基本イベントを 28,642 件獲得した．獲得された基本イベントは，例えば「{ お腹, 体 } ヲ 壊す」や「{ 雨, 雪 } ガ 降る」といったものであった．

3.2　蓋然的基本イベントペア候補の抽出

テキストに構文解析と談話関係解析を適用し，まず係り受け関係および蓋然的関係をもつイベントペアを抽出する．蓋然的関係は，イベント間の談話関係が談話標識によって明示されており，「原因・理由」または「条件」であるものとする．

本研究では，コーパスとして約 1.8 億文の日本語ウェブテキストを利用した．テキストからイベントペアを抽出するために，KNP[2]を用いた．KNP は係り受け

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

解析と，イベント間の明示的な談話関係の解析を行うツールである．KNP をテキストに適用した結果，約 2,000 万組の蓋然的イベントペア候補を抽出した．

続いて，解析結果から信頼性の高い部分を選択し，常識としての汎用的な内容を表すイベントペアに絞るため，以下の条件を満たすイベントペアを選択する．以下では，原因や理由を表す 1 つ目のイベントを前件イベント，結果などを表す 2 つ目のイベントを後件イベントと呼ぶ．

Reliable 前件イベントが後件イベントに曖昧性なく係る

1 文中に 3 つ以上の節がある場合，前件・後件イベント間の係り受け関係が解析誤りである可能性がある．このようなノイズを除くために，後件イベントが文末の節で前件イベントとの節間距離が 1 であるイベントペアを抽出する．

Basic 前件および後件イベントが両方とも基本イベントを含む

基本的には，前件・後件イベントが基本イベントを包含するかどうかをチェックするが，後件イベントの項は省略される可能性がある．この問題に対処するため，後件イベントが明示された項をもたない場合は，前件イベントから項を補完してイベントの包含関係を調べる．

例えば，「衝撃で窓が割れる → 取り替える」というイベントペアを考える．この場合，後件イベント「取り替える」に対して，前件イベントに含まれる項「窓」もしくは「衝撃」をもつ基本イベントが存在するかを調べる．この場合「窓を取り替える」という基本イベントが獲得されているため，このイベントペアは抽出される．

この条件を適用したところ，約 12 万組のイベントペアが選択された．

Basic 条件による効果を検証するため，Basic 条件を適用せずに抽出した蓋然的イベントペアおよび蓋然的基本イベントペアを無作為に 100 組ずつ抽出し，常識的に理解できるかを著者が評価した．その結果，蓋然的イベントペアは 47 組，蓋然的基本イベントペアは 77 組が常識をもとに理解できると評価された．従って，Basic 条件は常識レベルのイベント関係を抽出するのに効果的であることがわかる．

最後に，次ステップであるクラウドソーシングにおいてワーカーが的確に判断できるように，以下の後処理を行う．

- 事象性の薄いもの，ウェブ特有の機能表現を含むものを除くため，得られたイベントペアに含まれる基本イベントの頻度を計数し，高頻度の基本イベントを含むイベントペアを除く．高頻度の基本イベントは，「問題が無い」や「情報が満載」といったものであった．
- 指示代名詞や未知語を含むイベントペアを除く．

後処理の結果，51,841 組の蓋然的基本イベントペア候補を抽出した．

3.3 クラウドソーシングによる確認

蓋然的基本イベントペア候補から，クラウドソーシングを利用して蓋然的関係があるものを選択する．クラウドワーカーに，各イベントペアについて次の 2 択から選んでもらう．

1. A は B の原因・理由である
2. その他の関係，もしくは関係がない

ここでは，前件イベントを A，後件イベントを B と表現する．

各イベントペアを複数人のワーカーが評価し，過半数以上が一致した評価をそのイベントペアの評価とする．「A は B の原因・理由である」と評価されたイベントペアを蓋然的基本イベントペアとして採用する．

クラウドソーシングの結果，51,841 組から 34,170 組の蓋然的基本イベントペアを獲得した．

3.4 常識推論問題の生成

蓋然的基本イベントペアから常識推論問題を生成する．各問題の文脈は前件イベント，正解選択肢は後件イベントとする．誤り選択肢は，他のイベントペアデータ中の後件イベントから自動的に良質なものを選択する．

先行研究でも述べられているように，誤り選択肢をランダムに選択すると問題が簡単に解けてしまい，常識の学習に適さない [4]．正解選択肢に対して紛らわしい，良質な誤り選択肢を選択するために，以下の条件で選択を行う．

Similarity 正解選択肢との類似度が 0.5 より小さい．

選択肢間の類似度は，選択肢をベクトルで表現し，そのコサイン類似度で計る．選択肢のベクトル表現は，選択肢中に含まれる単語の平均ベクトルとする．

3 つの誤り選択肢は，類似度閾値条件を満たすものの内，正解選択肢との類似度が高い bin，正解

ご飯を食べたら， a. 早めに汚れを落とす b. 重低音のきいたモードもある c. 涼しげだ **d. さっさと歯を磨く**	音楽を聴くと **a. 気持ちが少しは和らぐ** b. 反動として身体に負担がかかる c. 魚は獲りにいける d. さっさと食って逃げるように店をでる	雨がひどかったので， **a. 延期になる** b. 2位争いになる c. ココのラーメンは大好きだ d. 早く札幌に着く

表 3: 作問結果の例 (太字は正解選択肢)

	学習	開発	テスト
問題数	27,304	3,278	3,288

表 4: データセットの統計

選択肢との類似度が中程度の bin，正解選択肢との類似度が低い bin から無作為に 1 つずつ選択する．

Length 正解選択肢に対して単語数の比が 0.5 から 2 までの範囲にある．

なお，誤り選択肢の候補が 3 つ以上得られない場合は問題を生成しない．

この結果，34,132 問の常識推論問題が生成された．表 3 に作問結果の一例を示す．

人間による解答精度を検証するため，生成した問題から無作為に 100 問サンプリングし，各問題に対して 5 人のクラウドワーカーから解答を収集した．多数決で決定した解答の精度は 92%であった．

作成した問題を 8 : 1 : 1 に分割し，それぞれ学習データ，開発データ，テストデータとする．作成したデータセットの統計を表 4 に示す．

4 計算機による解答実験

計算機による解答精度を検証するため，BERT による常識推論問題の解答実験を行った．BERT は，様々なタスクで高性能を達成した転移学習モデルである．本研究では，BERT のモデルとして，約 1,800 万文を含む日本語 Wikipedia コーパスで事前学習した BERT-LARGE モデルを用いた．

本実験においては，先行研究 [4] にならい，文脈と選択肢の組を特別な記号で区切って入力する．例えば，文脈が「お腹が空いたら」，選択肢が「ご飯を食べる」である場合，入力は “[CLS] お腹 が ... [SEP] ご飯 を ... [SEP]” となる．各 [CLS] トークンの中間層表現を線形層を通してロジットに変換し，最大値をとる選択肢を解答とする．

モデルの性能を Accuracy で評価した結果，BERT-LARGE の解答精度は 39%であった．高性能な転移学習モデルでも人間の常識推論能力との間には大きな隔たりがあり，バイアスが小さなデータセットになっていることが分かる．

また，条件 **Similarity** を除くと解答精度が 1%低下し，条件 **Length** を除くと解答精度が 2%向上した．この結果から，条件 **Similarity** は正解選択肢と酷似した誤り選択肢が混入して問題が解けなくなることを防ぎ，条件 **Length** は文長に関するバイアスを軽減していることが分かる．

5 おわりに

本論文では，コーパスからの自動抽出とクラウドソーシングにおける確認を組み合わせて，バイアスの少ない常識推論問題データセットを構築した．各問題は，基本的なイベント間の蓋然性を問う多肢選択式問題である．人間の解答精度は 92%と非常に高い一方で，計算機の解答精度は最高性能のモデルでも 39%と，常識推論能力に大きな隔たりがあることが判明した．現在，データセットの規模を拡大し，約 10 万問を含むデータセットを作成中であり，一般公開する予定である．今後は，構築したデータセットで常識を学習し，省略・照応解析や談話関係解析など他のタスクに応用することを検討する．

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．(公財) 日本漢字能力検定協会からの研究助成に感謝いたします．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, 2019.

[2] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation Artifacts in Natural Language Inference Data. In *NAACL-HLT*, 2018.

[3] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[4] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In *NAACL-HLT*, 2018.