文脈による危険度変化の予測のためのデータセット構築

勝又友輝[1]　竹下昌志[1]　ジェプカ・ラファウ[2]　荒木健治[2]
[1] 北海道大学大学院情報科学院
[2] 北海道大学大学院情報科学研究院
katsumata.yuki.v3@elms.hokudai.ac.jp
{takeshita.masashi,rzepka,araki}@ist.hokudai.ac.jp

概要

人工知能にも我々人間と同じような道徳的判断ができるシステムが必要である．しかし，現状の人工知能は，状況や文脈に応じた道徳的判断は困難である．例えば，「リング場でボクサーが人の顔を殴る」より「居酒屋でボクサーが人の顔を殴る」方が危険な行動であることを，人間は常識を持っているため簡単に判断できるが，人工知能には難しい．本研究ではこの問題を解決するため，文脈によって変化する人間の行動の危険度予測のためのデータセットを構築する．人手で動詞を元に人間の行動の文を作成し，7段階の危険度のアノテーションを行う．また，構築したデータセットを用いて，危険度予測の実験も行う．

1　はじめに

近年，人工知能の発展により，Siri や Pepper などのコミュニケーションエイジェントが登場し，我々の生活には人工知能が欠かせない存在となってきている．さらに，人工知能は履歴書の審査や融資の承認まで，様々な領域でますます権限を委ねられつつある．このような時代の変化に伴い，人工知能を有効に扱うためには，人間と同じように道徳的判断のできるシステムが必要である．

しかし，人工知能が道徳的判断をするにはまだ課題が多い．物事に対する道徳的判断は状況や文脈によって変動してしまう．人間は常識的な知識を持っているためこのような変化を捉えることができるが，人工知能はそのような常識を持っていない．そのため，状況や文脈などが変わった時に，人間と同じような判断をすることが難しい．人工知能への道徳的判断システムの導入には，文脈に依存した判断ができるようにすることが必要とされる．また，道徳的に正しいか否かという評価軸は不一致が起こりやすいという問題もある．人間でさえ意見が割れてしまうケースもあるため，人工知能が正しい判断をするのは困難である．

表 1　データセットの例，危険度スコアは安全（1）から危険（7）までの 7 段階とする

人間の行動の文	評価者 1	評価者 2	評価者 3
学生が音楽に合わせて手を叩く	1	2	1
学生が親の肩を強く叩く	5	5	6

本研究では，これらの背景から，常識的な知識による部分が大きく評価が一意に決まりやすい，人間の行動の危険度予測というタスクを提案する．このタスクでは，文脈が変化しても人工知能が人間と同等の危険度予測を行えることを目標とする．しかし，現状では，人間の行動の危険度を予測するための日本語のデータセットは存在しない．そこで本研究では，本タスクのためのデータセットを構築した．このデータセットは，**表 1** のように人間の行動の文とその行動の評価が 7 段階の危険度で構成されている．文脈が変化しても正しい判断を行えることを目指すため，動詞と主語は固定させ，動詞の目的語や文脈を追加する形で文を作成し，対象や状況が変化しても人間と同等の危険度予測ができるかを確認できるようにする．また，構築したデータセットを用いて，人間の行動の危険度予測の実験も行う．

なお，本研究での危険の定義は，「文が表す状況に登場する人物に，肉体的もしくは精神的に苦痛や傷害を与える可能性があること」とする．

2　関連研究

近年，人工知能に常識的な知識を持たせる研究や道徳的判断をさせる研究がさかんに行われている．

Maarten らは，常識的な推論に不可欠な推論知識をまとめた ATOMIC というデータセットを構築した [1]．ATOMIC は，「もしあるイベントが起こったら何が起こるか」という推論関係の知識で構成さ

れている．例えば，“if X pays Y a compliment, then Y will likely return the compliment” といった知識であり，このような推論関係を 87 万件まとめており，この種の知識グラフとしては最大のものである．

Rzepka らは，文脈を考慮した道徳的判断の研究を行った [2]．この研究では，日本語のコーパスで学習させた BERT モデル [3] を用いて「殴る」が文末に来るような文を生成した．そして，その文の表す行為が道徳的に許せるかどうかを人手で評価し，それを BERT が予測するという実験を行った．

しかし，Rzepka らの研究で生成されたデータセットには 3 つ問題がある．第一に，日本語 BERT モデルを用いて「殴る」という一つの動詞から生成された文は 50 文であり，データの量が不十分である．第二に，日本語 BERT モデルを用いて生成する方法では不自然な文が生成されることがある．第三に，道徳的に正しいという評価軸は評価者の中で不一致が起こりやすいという問題である．

本研究では，これらの問題の改善を目指し，道徳的判断という評価軸を改め，人間の行動の危険度予測というドメイン特化のタスクを提案する．また，現実的にあり得る人間の行動の文を得るため，クラウドソーシングを主に用いて，人間の行動の文を約 2 万文作成，危険度のアノテーションを行う．

3 データセット構築

本研究では，次の手順で，動詞を元に人間の行動の文を作成，および 7 段階の危険度のアノテーションを行う．

1. 関連研究から，危険な行動になり得る動詞のリストを作成する．（**3.1 節**）
2. 高頻度な述語項構造から，それぞれの動詞の主語を決定する．（**3.2 節**）
3. 主語と述語の組み合わせを基に対象語を人手にて穴埋めしてもらい，危険である例，危険でない例の文を作成する．（**3.3 節**）
4. 作成した文に，さらに文脈を付け加え，文の数を増加させる．（**3.4 節**）
5. 作成した文が表す人間の行動を，人手にて 7 段階で危険である可能性が高いか評価する．（**3.5 節**）

以降では，各手順の設定や方法について説明する．

3.1 動詞の決定

まず，危険な行動，危険でない行動，どちらにもなり得ると考えられる動詞のリストを作成する．動詞のリストの作成に当たって，2 つのデータを基にする．

1 つ目は，Alshehri らによる研究 [4] で定めた，物理的な危害を示すために使われるアラビア語の動詞のリストである．Alshehri らは，ソーシャルメディアにおける危険な発言の理解と検出のために，発言が危険か否かのラベルがついたデータセットの構築に取り組んだ．本研究では，Alshehri らがデータセット構築のために作成した 56 個の動詞（例：“kill”，“hit”）のリストを翻訳して活用する．

また，2 つ目は，日本語 Wikipedia エンティティベクトル [5][1] の頻出動詞である．日本語 Wikipedia エンティティベクトルとは，日本語版 Wikipedia の本文全文から学習した，単語，および Wikipedia で記事となっているエンティティの分散表現ベクトルである．また，頻出動詞の抽出には，日本語形態素解析システムである JUMAN[6] を使用する．

以上の 2 つのデータより，危険な行動，危険でない行動，どちらにもなり得ると著者が判断した動詞のみを抜粋し，25 個の動詞のリストを作成する．25 個のうち，Alshehri らの動詞のリストから 12 個，日本語 Wikipedia エンティティベクトルの頻出動詞から 13 個である．作成した動詞のリストを**表 2**，**3** に表す．

表 2 Alshehri らの動詞のリストから抽出した動詞

殺す	壊す	叩く	燃やす	切る	殴る
撃つ	打つ	刺す	割る	消す	与える

表 3 日本語 Wikipedia エンティティベクトルの頻出動詞のリストから抽出した動詞

投げる	落とす	蹴る	浴びる	奪う	使う	作る
飲む	食べる	歩く	扱う	運ぶ	当てる	

3.2 主語の決定

次に，動詞のリストを基に，それぞれの動詞の主語を決定する．主語の決定には，京都大学格フレームを用いた [7]．格フレームとは，用言とそれに関係する名詞を用言の各用法ごとに整理したものであり，京都大学格フレームは，Web テキストから自動構築した大規模格フレームである．これには，先ほ

1） http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/

ど定めた 25 個の動詞のそれぞれの頻出主語が記されている．これを基に，日本語 Wikipedia エンティティベクトルの人物を表す頻出名詞上位 50 件の中から，それぞれの動詞の頻出主語上位 5 件を選択する．ここで抽出対象を上位 50 件に絞った理由は主語の一般性を高めるためである．例えば，「カメラマンが____を切る」の対象語には「シャッター」や「電源」などの危険でない単語が入りやすく，危険な自然文と危険でない自然文の両方を獲得することは難しい．そこで，抽出対象を上位 50 件（例：「父」，「子供」）にすることで，このような問題の解決を試みる。

また、日本語 Wikipedia エンティティベクトルの名詞から抽出する名詞が人物か否かを判別するために，JUMAN を用いる．

以上より，それぞれの動詞に対して 5 件の主語を抽出した。例を挙げると，「殴る」という動詞は，「男」，「父」，「人」，「先生」，「女」という主語，「食べる」という動詞は，「人」，「子供」，「客」，「娘」，「自分」という主語となった．

3.3 対象語の穴埋め

主語と述語を基に，クラウドソーシングサービス[2)]を通じて，18 人のアノテーター（詳細は**付録 A** に示す）に現実的にありうる人間の行動の文の作成を依頼する．1 人あたり危険そうな例，危険ではなさそうな例をそれぞれ 125 文（動詞 25 個にそれぞれの頻出主語 5 個），合計 250 文の作成を依頼する．例えば，「子供が____を叩く」という文を与え，危険そうな例として「子供が友達の頬を叩く」，危険ではなさそうな例として「子供がタンバリンを叩く」といったような文を作成してもらう．

文作成依頼の際のルールは**付録 B** に載せる．結果，重複した例，ルールに合致しない例を除き，3,165 文の人間の行動の文が作成された．

3.4 文の拡張

作成した 3,165 文を元に，対象語の穴埋めと同様にクラウドソーシングサイトを通じて，文脈を増やし，現実的にありうる人間の行動の文の作成を依頼する．66 人のアノテーター（詳細は**付録 C** に示す）に 1 人あたり危険そうな例，危険ではなさそうな例をそれぞれ 150 文ずつ，合計 300 文の作成を依頼する．文作成依頼の際のルールは**付録 D** に載せる．アノテーションの結果，元の 3,165 文から 18,427 文が作成され、元の文と合わせて合計 21,592 文となった．

2) https://crowdworks.jp/

3.5 危険度のアノテーション

最後に，作成した文が表す人間の行動に対して，危険度のアノテーションを行う．文作成の際とは異なるアノテーターにより，7 段階で危険である可能性が高いか，安全である可能性が高いかのタグ付けを行う．作成した 21,592 文を 1 文につき 3 人ずつ，計 81 人のアノテーター（詳細は**付録 E** に示す）に，危険度のタグ付けを依頼する．また，アノテーションのスコアは，**図 1** ように安全（1）から危険（7）までの 7 段階とする．

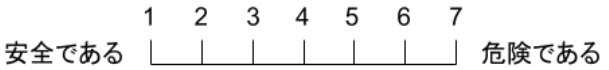

図 1 危険度アノテーションの評価スコア

アノテーションを行った結果，付けられた危険度スコアの分布は**表 4** のようになった．

表 4 評価の分布．評価の総数と文の総数（21,592 文）は一致しないが，これは 1 文につき 3 人ずつ評価を行ったためである．

危険度	1	2	3	4	5	6	7
評価の数	16,697	8,323	5,413	5,488	7,287	7,161	14,407

表 4 より，スコアが 3 または 4 である文が少ないという結果になった．これは，文作成の際に，危険そうな例，危険ではなさそうな例という形で，文を作成したため，どちらとも言えないような中間の値が少なくなったと考えられる．

また，文ごとに評価の標準偏差を求め，評価者によるばらつきがどれくらいかを確認する．文ごとに標準偏差を左から降順に整列させたグラフを**図 2** に示す．標準偏差の平均は 0.777 となった．1 未満という結果となり，全体的に評価のばらつきは大きくなく評価者の中で不一致があまり起きていないことが確認できた．しかし，**図 2** より，評価のばらつきが大きいものも存在したため，標準偏差が 2 以上のデータは信用できないデータとして除外した．結果，標準偏差の平均は 0.692 となり，より信頼できるデータとして予測実験に使用する人間の行動の文とその行動の危険度の組み合わせは 20,453 件となった．

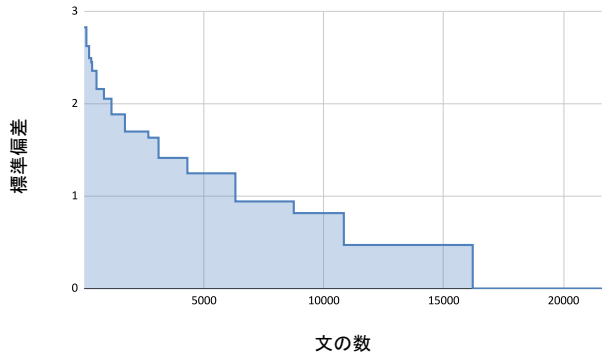

図 2　文ごとに標準偏差を計算し，左から降順に整列させたグラフ．

4　危険度予測の実験

機械学習による危険度予測実験の精度を検証するため，LSTM, Bidirectional LSTM（以下，BiLSTM とする）, BERT の 3 つのモデルによる回帰分析を行う．LSTM と BiLSTM のエンベディングには，One-hot 表現と日本語 Wikipedia エンティティベクトルを使用する．BERT のモデルとしては，約 1,800 万文を含む日本語 Wikipedia コーパスで事前学習された BERT-base モデル[3]を用いる．

危険度スコアの正解値は 3 人の評価者の中央値，平均値，2 つの場合で実験する．それぞれ構築したデータセットを評価値ごとに層化抽出法を用いて 8:1:1 に分割し，学習データ，開発データ，テストデータとした．データセットの詳細は**付録 F** に示す．

また，回帰分析の評価には，一般に広く用いられている 2 つの精度評価指標[4]，平均絶対誤差（Mean Absolute Error，以下，MAE とする）と二乗平均平方根誤差（Root Mean Squared Error，以下，RMSE とする）を用いる．

MAE とは，正解値と予測値との誤差の絶対値を平均したものである．これは，データの数を n，正解値を y，予測値を f としたとき，**式 (1)** のように表せる．

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{1}$$

次に，RMSE とは，正解値と予測値との誤差の二乗を平均して平方根をとったものである．これは，MAE より正解値と予測値との誤差が大きい例の影響を受けやすいという特徴がある．**式 (2)** のように表せる．

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2} \tag{2}$$

3）https://nlp.ist.i.kyoto-u.ac.jp/index.php?ku_bert_japanese

4）https://scikit-learn.org/stable/modules/model_evaluation

5　実験結果

表 5，**6**，**7** に実験結果，例を示す．

表 5　危険度の正解値を評価の中央値にした場合の実験結果，最も高い精度を太字で示す．

	LSTM		BiLSTM		BERT
	one-hot	Word2vec	one-hot	Word2vec	
MAE	1.232	1.421	1.221	1.278	**0.899**
RMSE	1.775	1.809	1.757	1.658	**1.329**

表 6　危険度の正解値を評価の平均値にした場合の実験結果，最も高い精度を太字で示す．

	LSTM		BiLSTM		BERT
	one-hot	Word2vec	one-hot	Word2vec	
MAE	1.050	1.174	1.041	1.085	**0.749**
RMSE	1.519	1.507	1.485	1.472	**1.082**

表 7　危険度の正解値を評価の中央値にした場合の BERT の実験結果の例

人間の行動の文	正解値	予測値
男が体験レッスンでボクサーを殴る	2	2.096
男が酔っ払ってボクサーを殴る	6	5.992

表 5，**6** より，全体では BERT が最も良い精度となった．**表 7** の例では，文脈の違いによる危険度変化の正確な予測がされた．

6　おわりに

本研究では，文脈によって変化する人間の行動の危険度予測のためのデータセットを構築した．人間の行動の文とその行動の危険度の組み合わせが 20,453 件となった．また，機械学習を用いた回帰分析による予測を行った．実験の結果，BERT を用いた予測が最も精度が高いことが明らかとなった．

今後は，予測の精度を上げるため，他のモデルとの比較検討を行う必要がある．

また，データセットについても，さらなる検討をすべきである．文が表す人間の行動から，連想される状況や行動を踏まえた予測の検討も重要であると考える．ある行動を基に次に起きる事象を予測することは，文脈の考慮にも繋がり，予測の精度を上げることに貢献できると考えられる．そのため，因果関係の知識やシナリオを追加して，データセットを拡張することを今後検討していきたい．

謝辞

本研究は JSPS 科研費 17K00295 の助成を受けたものです.

参考文献

[1] 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. **AAAI**, 2019.

[2] Rafal Rzepka, Sho Takishita, and Kenji Araki. Bacteria lingualis on bertoids - concept expansion for cognitive architectures. **Technical Report of JSAI Special Interest Group for Artificial General Intelligence**, 2020.

[3] Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, and Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, 2019.

[4] Ali Alshehri, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. Understanding and detecting dangerous speech in social media. **Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools**, 2020.

[5] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会 (NLP2016), 2016.

[6] Hajime Morita, Daisuke Kawahara, and Sadao Kurohashi. Morphological analysis for unsegmented languages using recurrent neural network language model. **Proceedings of EMNLP 2015: Conference on Empirical Methods in Natural Language Processing**, 2015.

[7] 河原大輔, 黒橋禎夫. 黒橋禎夫：高性能計算環境を用いた web からの大規模格フレーム構築. 情報処理学会 自然言語処理研究会, 2006.

A 3.3 節の対象語の穴埋めのアノテーター情報

20 代から 60 代の男性 7 名，女性 11 名

B 3.3 節の対象語の穴埋めのルール

- 埋める部分は，物や人を問わず，実体のあるもの（非抽象的対象）とし，また代名詞（例：「私」，「彼」，「彼女」）は使ってはいけない．
- 「修飾語＋名詞の形」でも良い．（例：「母親の肩」，「腐ったみかん」，「汚い水」）
- 「名詞句＋に＋名詞句」，「名詞句＋から＋名詞句」の形はできるだけ避け，目的語が一つになるようにする．
- 主語と同じ意味のものを使ってはいけない．（例：「女性が【女】を殴る」）

C 3.4 節の文拡張のアノテーター情報

20 代から 60 代の男性 21 名，女性 45 名

D 3.4 節の対象語の穴埋めのルール

- 元の文と全く同じ文は書いてはいけない．
- 元の文の単語は変えずに，文脈を増やして文を作成する．
- 文の終わりに文脈を増やしてはいけない
- 危険そうな例，危険ではなさそうな例のどちらかが，どうしても思いつかない場合は空欄にする．

E 3.5 節の危険度評価のアノテーター情報

20 代から 60 代の男性 38 名，女性 43 名

F 実験に用いたデータセットの詳細

表 8 評価の中央値の分布

危険度	文の数
1	5226
2	2813
3	1546
4	1560
5	2408
6	2570
7	4330

表 9 評価の平均値の分布

危険度	文の数
1 以上 2 未満	5528
2 以上 3 未満	2772
3 以上 4 未満	1839
4 以上 5 未満	1882
5 以上 6 未満	2927
6 以上 7 未満	3568
7	1937

表 10 実験に用いたデータセットの統計

	学習	開発	テスト
文の数（中央値の場合）	16,360	2,046	2,047
文の数（平均値の場合）	16,359	2,046	2,048