日本語大規模データセットにおける嘘つき検出

青木洋繁
静岡大学 情報学部
haoki@kanolab.net

狩野芳伸
静岡大学 情報学部
kano@inf.shizuoka.ac.jp

概要

言語情報のみでの欺瞞・嘘つき検知[1)]は人間にとって難しい. 日本語での対話データを用いた欺瞞・嘘つき検出の研究はこれまでいくつかあったが, 非言語情報を主に用いたものが多く, 言語情報のみを用いたものが筆者の知る限り存在しなかった. また, 英語で行われた欺瞞・嘘つき検出の研究には言語情報のみの対話データを用いたものがあったが, 欺瞞・嘘つきラベルが付けられたデータセットは集めにくいという特性上, これまで行われてきた研究で使用されていたデータセットは規模が小さかった. そこで本研究では, 日本語の大規模データセットを用い, 英語の先行研究の手法が日本語のデータセットにおいても有効かどうかの検証に加え, 深層学習モデルを学習させ, 英語の先行研究の手法と比較した. Accuracy 約 72.4%で分類に成功し, 先行研究よりも安定して嘘つきを検出できた.

1 はじめに

言語情報のみの欺瞞・嘘つき検知は人間にとって難しく, 人間はオンラインのテキストコミュニケーションで, 相手を信頼する傾向があることが分かっている [1]. SNS の流行やコロナウイルスの影響でオンラインでのコミュニケーションの機会が増えた今, 言語情報のみの対話データでの嘘つき検出は重要であり, 客観的に対話相手が嘘つきかどうかを判定出来るモデルは有用である.

英語の書き言葉のみの対話データを用いた欺瞞・嘘つき検出の研究はいくつかある [2], [3]. しかし, 日本語の対話データでの欺瞞検知の研究では非言語情報も用いたものが主であった [4], [5], [6]. また, これまでの対話データを用いた嘘つき検出の研究では, 嘘付きラベルが付与されたデータの取得が難しいという特性上, 小規模なデータが用いられていた.

本研究では, インターネット上で行われていた人狼ゲームの人狼 BBS[2)]のログを使い, 大きなデータセットで深層学習モデルを学習させ, 嘘つきを予測できるモデルの作成を目指す. また, モデルが予測の際に着目した箇所を可視化し, どのように嘘つきの特徴を捉えているかを検証する.

1.1 人狼ゲームとは

人狼ゲームとは, プレイヤーを騙す人狼陣営と嘘付きを見破る市民陣営にランダムに分け, それぞれの勝利のために対戦するゲームである. 人狼は夜になると人間を一人襲撃する. 人狼による襲撃を止めるため, 昼間に村人全員で話し合いを行い, 一日に一度, 最も怪しい村人を投票で 1 人選び処刑する. 市民陣営は人狼を全て処刑すれば勝利し, 人狼陣営は市民陣営よりも多くなった時点で勝利する. 市民陣営には他のプレイヤーの役職を知る事が出来る占い師などの能力者がおり, 市民陣営の勝利に大きく貢献する. ただし, 役職はランダムに割り振られ, ゲームが終わるまで正解は明かされないため, 能力者と人狼以外 (人狼同士はお互いを認識している) は, 他のプレイヤーの役職を知る事が出来ない. したがって, 人狼陣営は容易に市民陣営を欺くことができ, 市民陣営が勝つためには発言を吟味しながら, 嘘つきを見破る必要がある.

2 関連研究

2.1 言語心理学の特徴量による嘘つき検出

Girlea ら (2016)[2] は, Barnwell によって収集された英語版人狼ゲームのログ (86 ゲーム分)[3)]を用い, これまで言語心理学で提唱されてきた嘘つきの特徴を特徴量として, ロジスティック回帰や Random

1) 本研究では, 一発話による「欺瞞」の検出ではなく, 複数の発話からの「欺瞞を行う人」検出すなわち「嘘つき」検出を扱った.

2) 人狼 BBS のログ: http://ninjinix.com/

3) 先行研究で用いられた人狼ゲームのログ: https://www.brenbarn.net/werewolf/logindex.html

Forest 等で騙し役をプレイヤーの全発話単位で予測し, Random Forest で 90.87%の精度で嘘つきを予測することが出来た. また, logistic 回帰の結果から英語の嘘つき検出において特に TTR(type token ratio) が予測に寄与することを示した. ただし, この実験で用いられたデータは小規模であり (701 インスタンス, 内 116 インスタンスが騙し役), 言語は英語であった. 今回はこの手法を日本語の大規模データセット (人狼 BBS データセット) でも同様に有効か検証し, 深層学習モデルと比較した.

2.2 深層学習モデルを利用した欺瞞検知

Peskov(2020) ら [3] は, Diplomacy という国盗りゲームを実験参加者にプレイさせ, 発話に送り手が嘘を付いたかのアノテーションと受け手が嘘の発話に見えたかどうかの双方向のアノテーションを付けたデータセットを作成した. このデータセットを使い, ロジスティック回帰とニューラルモデルで発話単位での真偽の予測を行った. 結果, 人間のベースラインと近い精度が得られた. ただし, この研究は, 発話の真偽予測を対象としているため, 対話データ全てを使い騙し役を予測する本研究と直接比較することは出来ない.

3 データセット

3.1 人狼 BBS データセット

今回はデータセットに人狼 BBS のログデータを用いる. 人狼 BBS は電子掲示板上で以前行われていた人狼対戦ゲームで, インターネット上に過去 10 年分のログデータ (7,264 ゲーム分) が公開されている. 初日に襲撃の対象になるコンピュータを含め, 10 人から 16 人の参加者で 1 ゲーム当たり 1 週間程度で行われていた. 一日に一度更新があり, そのタイミングで投票が行われ, 最も怪しい村人の処刑, 人狼の襲撃, 能力者の能力使用が行われる. 役職の役割等, 人狼 BBS についての更に詳しい情報は, まとめサイト[4] を参照されたい.

人狼 BBS では, 人狼同士が昼間の話し合いの間も他の村人に気付かれることなく会話をすることができるため, 人狼同士連携の取れた高度な嘘が展開されていた. 図 1 に人狼 BBS のログの一部を示した. この例では, ジムソンが市民陣営で, カタリナが人狼陣営である.

4) 人狼 BBS まとめサイト https://wolfbbs.jp/

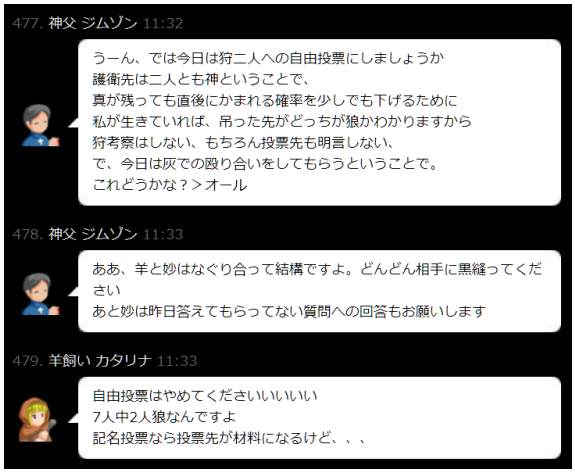

図 1 人狼 BBS のログの例

図 1 の吹き出し 1 つを 1 発話として, プレイヤーの 1 ゲーム内の全発話を 1 インスタンスとした. 今回はは人狼 BBS の役職の内, 人狼陣営の人狼, 狂人と市民陣営の村人, 占い師, 霊能者, 狩人を使用した. 本研究では, 市民陣営を欺き自陣営の勝利を得ようとする人狼陣営のプレイヤーを「嘘つき」とみなし二値分類を試みた. 前処理を行った後, 市民陣営を 0, 人狼陣営を 1 とラベル付けしたところ, 市民陣営 61,839 インスタンス, 人狼陣営 26,977 インスタンスとなった.

3.2 前処理

web からクロールした人狼 BBS データセットから, 不要な文字列 (顔文字, 絵文字, URL, レスアンカー) の削除, 表記ゆれの統一 (半角ひらがな, 半角アルファベットに統一), 空白・改行の削除, 数字の 0 への変換, モデルが人狼用語 (キャラクター名) に頼って予測をしないようにキャラクター名やニックネームの<person>トークンへの置換, 嘘つき検出に関係のない部分 (独り言, 人狼同士の対話, 墓地での会話, プロローグ, モノローグ, コンピュータの発話) の削除, 10 発話以下のプレイヤーの削除を行った.

尚, キャラクター名やニックネームは, 人狼 BBS まとめサイトの登場人物のページ[5] を参考にしてリストアップし, ログデータを見て手作業で更に追加した. そして, 正規表現を用い, 関係ない箇所が置換されないように工夫して置換した.

人狼 BBS を何度もプレイしているユーザの特徴からモデルが予測をしてしまう事を避けるため, 同じユーザが train:valid:test に跨って出現しないよう

5) 人狼 BBS まとめサイト 登場人物: https://wolfbbs.jp/%C5%D0%BE%EC%BF%CD%CA%AA.html

にデータセットを train:valid:test を 8:1:1 に分割した. 分割後の市民陣営と人狼陣営の数はそれぞれ, train(49,601 : 21,447), valid(6,138 : 2,751)　test(6,100 : 2,779) となった. ここから人狼陣営をオーバーサンプリングして, 市民陣営と人狼陣営が 1:1 になるようにした. オーバーサンプリング後の統計量を表 1 に示した.

前処理を行ったデータセットから, mecab(ipadic neologd[7]) を使い形態素分割した後, 先行研究 [2] で用いられた言語心理学の特徴量をカウントしたデータセット, プレイヤー毎に発話をまとめたデータセットを作成した.

表 1　オーバーサンプリング後の統計量

	train	valid	test
プレイヤー数	99,202	12,276	12,200
異なりユーザ数	13,624	1,660	1,942
合計発話数	7,119,118	878,121	870,219
平均発話数	71.76	71.53	71.33
最大発話数	199	180	178
最小発話数	11	11	11

4　モデル

4.1　Random Forest

先行研究 [2] の言語心理学の特徴量を用いたモデルの中で, Random Forest が最も精度が高かったため, 本研究では Random Forest を使って先行研究を再現した. 先行研究 [2] で使用された特徴量の内, 認知的な複雑さを測るために用いられていた"6 文字より長い単語"は除外し, 英語の前置詞の数の代わりに助詞の数を用いた. また, ネガティブ感情語数は東北大の極性辞書を用いてネガティブ感情語をカウントしたものを用いた. 和らげ・ぼかし表現, 自己参照, 認知語, 動作語, 除外語, 否定語, 動作語, 知覚語は mecab の辞書にあるか確認しながらリストを作り, 発話を形態素解析した後, リストに存在する語をカウントした. その他の品詞の数は, mecab の形態素解析結果に応じてそれぞれカウントした. 以下に使用した特徴量をまとめた.

- TTR (type-token ratio)
- 口ごもり語の数 (hesitation)(えーと, うーん)(14 語)
- ネガティブ感情語数
- 形態素数
- 自己参照の数 (私, 僕, おいどん)(45 語)
- 否定語の数 (しない, ない, 違う)(12 語)
- 和らげ・ぼかし表現の数 (hedge)(みたいな, 多分, ちょっと)(34 語)
- 認知語の数 (信じる, 考える, 思い出す)(50 語)
- 動作語の数 (到着する, 走る, 歩く)(31 語)
- 知覚語の数 (触る, 見る, 聞く)(30 語)
- 除外語の数 (以外, その他)(7 語)
- 助詞と接続詞の数
- 代名詞の数
- 形容詞の数
- 名詞の数
- 動詞の数
- 接続詞の数
- 助詞の数

4.2　HAN(Hierarchical Attention Network)

HAN(Hierarchical Attention Network)[8] は, attention 機構を擁した 2 層の bidirectional GRU からなるモデルである. 1 層目は単語レベルで 2 層目は文レベルである. 2 層目は, 1 層目の単語の重み付き和を入力として受け取る (図 2). つまり, 一発話ずつ読み, これまでの発話と照らし合わせながらそのプレイヤーが人狼かどうかを判断する. この処理は, 人間が人狼をプレイするときの処理に似ている. また, 人間が着目してる箇所と深層学習モデルが着目している箇所を定量的比較分析した研究で bidirectional LSTM の注目している箇所は人間に近いということが報告されている [9]. 予測の結果だけでなく, 人間に分かりやすい根拠を attention の可視化等を用いて示せば, 実際に人間が嘘つきを検出するときの助けになると考え, HAN を選択した.

今回, training データと前処理で除外して学習・評価に使用しなかったデータ (プロローグ等) から学習した Sentencepiece[10](bpe, skip-gram, 語彙数 32,000) を用いて, 発話をサブワードに分割した. サブワードの分散表現の獲得には, fasttext[11](200 次元) を用いた.

5　実験

実験の評価は, accuracy, precison, recall, f1 で行った. 実験の結果, HAN で Acuracy 約 72.5%で分類に成功し, 先行研究の手法を上回った (表 2, 3). 先行研究の手法の accuracy が約 52.4%に落ちたのは, 英語のデータセットで有効であった特徴量が日本語のデー

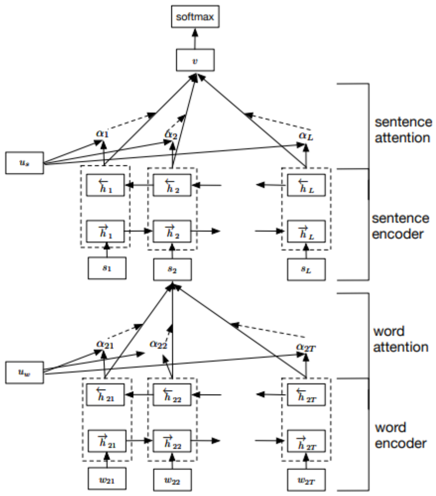

図 2　Hierarchical Attention Network

タセットには適していなかったこと, データセットのサイズが大きく変わったこと (先行研究のデータセット: 701 インスタンス, 人狼 BBS データセット: 88,816 インスタンス) が考えられる.

表 2　実験結果

モデル	accuracy	precision	recall	F1
Random Forest	0.524	0.560	0.524	0.439
HAN	0.725	0.726	0.725	0.725

表 3　陣営毎の precision, recall, f1, support(HAN)

陣営	precision	recall	f1	support
市民陣営	0.73	0.71	0.72	6,100
人狼陣営	0.72	0.74	0.73	6,100

6　モデルが着目した箇所の分析

モデルが重視した箇所を検証するために, モデルが 90%以上の確信度で予測した時, 文の中で attention の重みが一定以上 (発話中のサブワード全体に均等に割り振った確率以上[6]) かかっていた trigram の重みを足し合わせ, 重みが大きい順に True Positive(TP), False Positive(FP), True Negative(TN), False Negative(FP) に分けて, 重みの和が大きい順に並べた. それから, TP と TN 特有の trigram を TP-(TN+FP+FN), TN-(TP+FP+FN) のように差を取り, wordcloud で可視化した. wordcloud を見ると, 90%以上の確信度でモデルが人狼陣営だと判定し, 実際に正解だった時, モデルが着目した箇所は曖昧な当たり障りのない箇所が多かった (図 3). 一方, 90%以上の確信度でモデルが市民陣営だと判定し, 実際に正解だった時, モデルが着目した箇所には<person>トークン (キャラクター名やニックネーム) が列挙されている箇所が多く (図 4), <person>を列挙するのは, 他のプレイヤーの役職推定をするときに多いため, プレイヤーが具体的に自分の考えを表明した箇所にモデルが注目していたのだと考えられる. このように処理が人間に近い階層型モデルが確信をもって予測して正解した時には, 人間にも分かりやすい基準 (無難な発話ばかりするプレイヤーは人狼) を持って予測をする傾向がある事が分かった. 更なる可視化の例は付録に示した.

6)　trigram の threshold の求めかた

$$threshold = \frac{3}{\text{発話長} - 2} \quad (1)$$

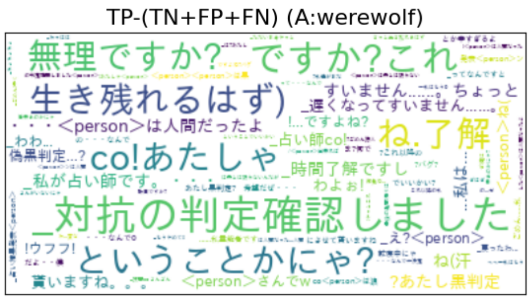

図 3　モデルの確信度 90%の時の TP 特有の trigram

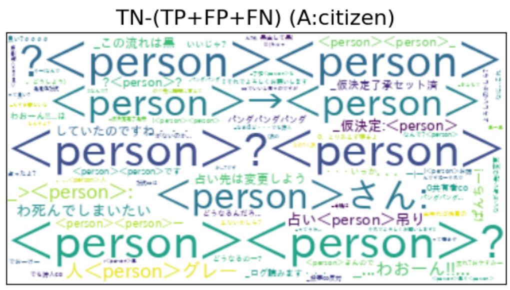

図 4　モデルの確信度 90%の時の TN 特有の trigram

7　おわりに

本研究では, 日本語の大規模データセットを用い, 深層学習モデルを学習させ, 先行研究のモデルよりも高い精度で嘘つきを検出できるモデルを作成した. さらに, モデルが予測した時に着目した箇所を可視化すると, 人間にも解釈しやすい基準で分類している事が分かった. 今後はこのモデルが一般的な嘘つきの特徴を捉えられているかを検証するため, 詐欺等のデータでこのモデルが嘘を見抜けるかどうかを検証したい. また、bidirectional GRU を BERT や RoBERTa, Sentence BERT に変更するなど, モデルの構造を工夫して, HAN に捉えられなかった嘘を捉えられるモデルの構築を目指したい.

謝辞

本研究を行うにあたり, 人狼 BBS のデータ使用を許可していただいた ninjin 氏に心より感謝致します.

参考文献

[1] Ben Shneiderman. Designing trust into online experiences. **Commun. ACM**, Vol. 43, No. 12, p. 57–59, dec 2000.

[2] Codruta Girlea, Roxana Girju, and Eyal Amir. Psycholinguistic features for deceptive role detection in werewolf. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 417–422, San Diego, California, June 2016. Association for Computational Linguistics.

[3] Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. It takes two to lie: One to lie, and one to listen. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3811–3854, Online, July 2020. Association for Computational Linguistics.

[4] Yuiko Tsunomori, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. Comparison of effective features and analysis of questions towards dialogue-based deception detection.

[5] Shohei Takabatake and Kazutaka Shimada. Construction and analysis of liar corpus. http://www.pluto.ai.kyutech.ac.jp/~shimada/paper/HCG2017_takabatake.pdf. Accessed: 2022-1-5.

[6] Kenichiro Tsuji, Sho Mitarai, and Nagisa Munekata. Investigation and analysis of utterance information based on the suspects' actual voice of the communications fraud. **Proceedings of the Annual Conference of JSAI**, Vol. JSAI2021, pp. 3J1GS6a01–3J1GS6a01, 2021.

[7] Sato Toshinori. Neologism dictionary based on the language resources on the web for mecab, 2015.

[8] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 1480–1489, San Diego, California, June 2016. Association for Computational Linguistics.

[9] Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4596–4608, Online, July 2020. Association for Computational Linguistics.

[10] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. **CoRR**, Vol. abs/1808.06226, , 2018.

[11] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. **arXiv preprint arXiv:1607.04606**, 2016.

付録 A

A.1 可視化例

以下に unigram での可視化例を示した (図 5, 6). サブワード・文の重みが threshold よりも大きいときに色付けされた. unigram での threshold は (2) で決定した.

$$threshold = \frac{1}{\text{発話長}} \tag{2}$$

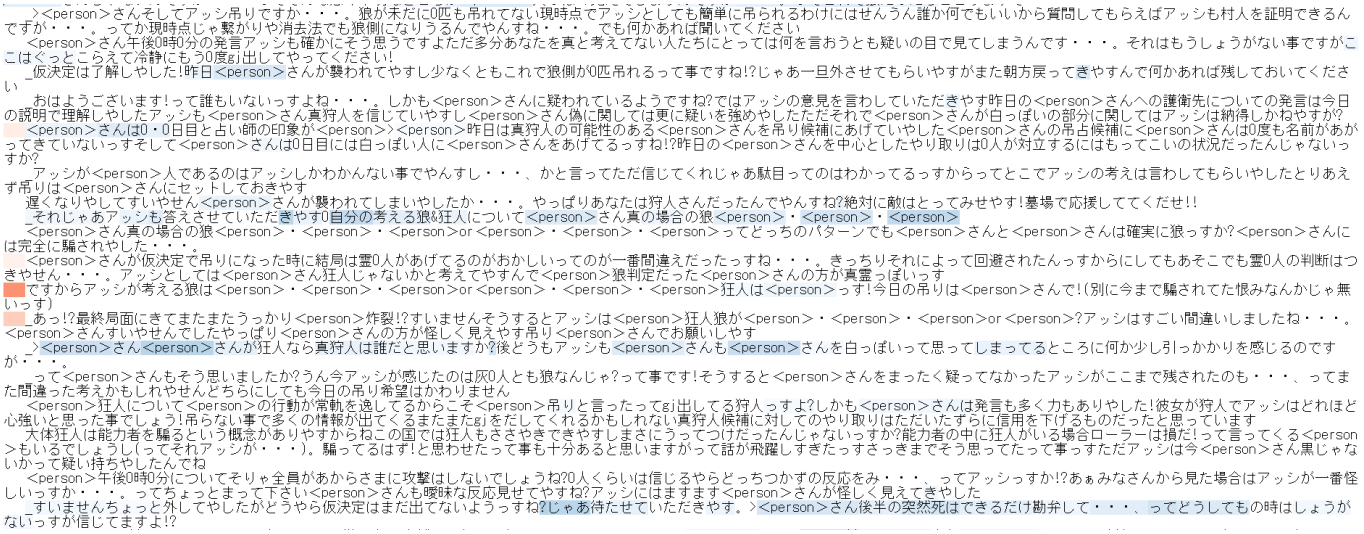

図 5 TP 確信度 92.81%での attention 可視化の例

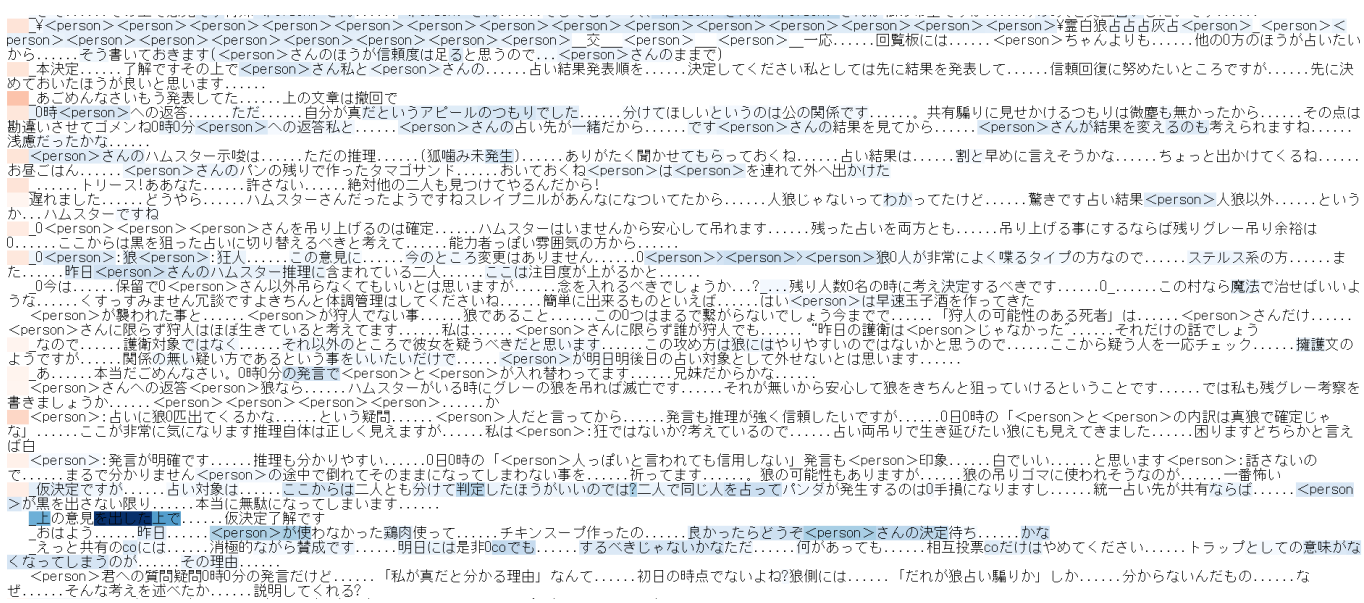

図 6 TN 確信度 99.62%での attention 可視化の例