機密情報検知における生成 AI を用いたデータ拡張

岸波洋介 [1]　藤井諒 [1]　加藤善大 [1]
[1] フューチャー株式会社
{y.kishinami.rh, r.fujii.6d, y.kato.v7}@future.co.jp

概要

企業において，電子メールやチャットの誤送信等，様々な場面に機密情報漏えいのリスクが存在する．これに対して，入力文中の機密情報を事前に検知することが有用であると考えられるが，モデルの学習には大規模なデータセットが必要となる．そこで本研究では，生成 AI を用いた機密情報検知のデータ拡張手法を提案する．実験では提案手法で作成したデータを用いて企業名，および法人名を検知するモデルの学習を行い，有効性を調査した．実験の結果，生成 AI を用いてエンティティを拡張する手法によりベースラインに比べ F1 で 2.5%の大幅な精度向上を確認した．

1　はじめに

業務のデジタル化に伴い情報管理に関するセキュリティリスクが高まっている．特に，個人や顧客に関する機密情報の漏えいは重大な問題である．電子メールやチャット，SNS の誤送信に加えて，最近では ChatGPT をはじめとした生成 AI を活用する企業も増えており，情報漏えいのリスクは様々な場面に潜んでいる．情報漏えいを防ぐにあたり，重要なのは機密情報の送信を未然に防ぐことである．そのための一つのアプローチとして，入力文に含まれる機密情報を検知し，送信前にユーザに通知することが考えられる．これは自然言語処理分野では従来より固有表現抽出 (NER) の文脈で取り組まれてきた．

近年，固有表現抽出では BERT [1] 等の事前学習済み言語モデルを固有表現ラベル付きデータセットでファインチューニングする手法が主流である．しかしながら，問題設定によってラベル定義が異なることや，アノテータの確保等の観点から高品質なラベル付きデータセットを大規模に構築することは高コストである．

一般に，このような課題への対処としては，既存のデータセットを元にしたデータ拡張が行われる [2, 3]．固有表現抽出においても，単純なルールに基づく手法 [4, 5] や，言語モデルを利用し文の一部や全体を言い換える手法 [6, 7] 等様々な手法でデータ拡張が行われている．ルールに基づく手法では，学習データセットの他の事例のエンティティを用いる，同一文内で順序を入れ替える等，限られた情報源を元にデータ拡張が行われることが多い．そのため，学習データの多様性が向上しないことが指摘されている [8]．これに対して，言語モデルを利用した手法では，言語モデルの表現力の範囲で多様性を大幅に向上できる可能性がある．既存研究では元の文と意味的な一貫性を保ったまま言い換えを行う等の手法の有効性が示されている [7]．しかしながら，企業や人物等に関する事実は今後も永続的に増えていくものであり，現時点で事実に反する記述も将来に渡って誤りであるとは言い切れない．したがって，現時点で事実として正しいことや，元の文と意味的に一貫していることは必ずしも重要ではないと考えられる．

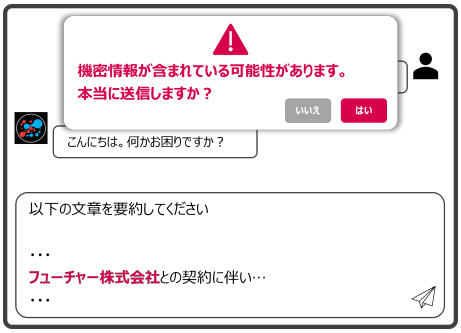

図 1　機密情報検知の概要．

そこで本研究では，生成 AI を用いた事実性に捉われない多様なデータ拡張を提案する．実験では，生成 AI を用いてエンティティを拡張する手法により，ベースラインに比べ F1 で 2.5%の大幅な精度向上を確認した．また，分析を通して，拡張したデータで学習することが文字種の複合を含む複雑な企業・法人名への頑健性を向上することが示唆された．

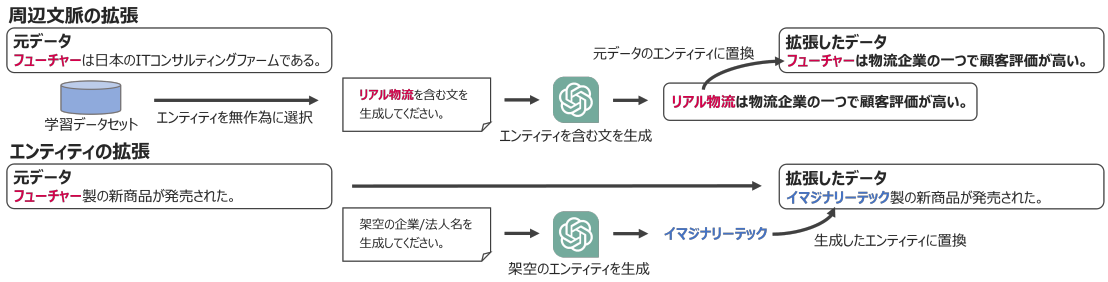

図 2　生成 AI を用いたデータ拡張の概要.

2　関連研究

大規模なニューラルネットワークの学習は一般に多量のデータを要する．そのような背景において，高品質なデータを人手で大量に作成することは大きなコストを伴うため，データ拡張と呼ばれる手法が活発に研究されている [2, 3]．自然言語処理におけるデータ拡張は大きく，文の一部を交換・挿入・削除するといったルールベース [9, 10]，逆翻訳 [11] やマスク言語モデルによるトークン予測 [12] といったモデルベースの手法に二分される．固有表現抽出においても，かねてより多くのルールベース，モデルベースのデータ拡張手法が提案されてきた．

Dai ら [4] は，文中のエンティティをデータセット中の同じラベルのエンティティに置き換える Mention Replacement (MR) 等，合計 4 つのルールベースによるデータ拡張を提案している．Dai らの調査では，これらの単純なデータ拡張が低資源な固有表現抽出で極めて効果的であることが示唆された．寺本ら [5] は，Dai らの手法で非文が生成される可能性があることに着目し，POS タグによる制約を課すことで非文の生成を抑える手法を提案している．しかしながら，これらの既存手法は限られた情報源を元に置換等の操作を行うため，学習データの多様性を向上させることができない [8]．

一方で，学習済み言語モデルを利用したモデルベースの手法も多く研究されている [6, 7, 8, 13, 14]．Li ら [6] は，系列変換モデル [15] を使用してエンティティ以外の文脈を拡張する手法を提案している．Sharma ら [7] は，GPT-3 [16] をはじめとした複数の言語モデルを用いて言い換え文を生成し，高品質な言語モデルによる言い換え文の生成が汎化性能の向上に寄与することを示した．しかしながら，Sharma らの手法では元の文との意味的な一貫性が重要視されており，多様性の向上は限定される．

本研究では，文として自然な構造を保っていれば，一見して事実に反すると思えるような文でもモデルの学習に有用であると仮定し，多様なエンティティ・文脈のデータを生成することに着目する．

3　タスク: 機密情報検知

固有表現抽出タスクで一般的に使用される固有表現は必ずしも機密情報ではない．例えば IREX [17] で定義される固有表現のうち，日付表現や時刻表現に起因してセキュリティ的な問題が発生する可能性は低い．一方で，パスワード等固有表現として定義されない表現でも検知すべきものは存在する．そのため，機密情報の事前検知を目的とした場合，一般的な固有表現抽出のラベル定義は適していないと考えられる．企業における機密情報として，第一に挙げられるのは社内や顧客の情報である．したがって本研究では，企業・法人名に着目し，与えられた文から企業・法人名を検知するタスクに取り組む．

4　生成 AI を用いたデータ拡張

本節では提案するデータ拡張手法について述べる．本研究では，生成 AI を用いたデータ拡張手法として大きく二つの手法を提案する．

4.1　周辺文脈の拡張

「フューチャーは日本の IT コンサルティングファームである。」という文から「フューチャーは物流企業の一つで顧客評価が高い。」という文を生成することを考える．これは一見すると事実に反するが，機密情報検知のタスクにおいては本文中の検知対象表現が必ずしも特定のエンティティと紐づいている必要はない．つまり，「フューチャー」は実在する IT コンサルティングファームでも，将来的に誕生する可能性のある物流企業でも構わない．そこで本研究では，生成 AI の「ハルシネーション」を逆手に取り，エンティティを含む文を自由に生成させることでデータ拡張を行う．具体的には，まず

表 1　作成したデータの例．太字は検知対象の企業・法人名を示す．

元データ	周辺文脈の拡張	エンティティの拡張
東洋パルプ株式会社 は、かつて存在した製紙会社である。	新しい都市計画に基づき、**東洋パルプ株式会社** は持続可能な都市の建設を推進しています。	**ユートピア食品** は、かつて存在した製紙会社である。
キングレコード で発売されたシングル曲は、すべてシングルバージョンで収録されている。	**キングレコード** は、輸送業界で革新的な技術を展開し、国際的に高く評価されています。	**ジャドラン医療機器** で発売されたシングル曲は、すべてシングルバージョンで収録されている。

ある文に対して，文内のエンティティとは独立に学習データセット中のエンティティを選択し，該当のエンティティを含む文を生成 AI を用いて生成する．その後，生成した文に含まれるエンティティを元の文のエンティティに置換する (図 2)．ここで，生成 AI は稀に文法誤りを含む文や非文を生成してしまう場合がある．そのため，生成された文を再度生成 AI に入力し，文法誤りを訂正させることで，低品質な学習データが生成されることを抑制する．

4.2　エンティティの拡張

「フューチャー製の新商品が発売された。」という文を考える．このとき，機密情報検知のタスクを解く上では，新商品を発売した企業が「フューチャー」であるという情報は本質ではない．現に，エンティティ置換の有効性は Dai ら [4] の提案する MR によって示されている．しかし，MR ではデータセットの他の事例に含まれるエンティティを用いるため，データセット全体のエンティティの多様性は向上しない．また，Sharma ら [7] は MR で使用するエンティティを GPT-3 を用いて生成しているが，意味的な一貫性を保持するために元のエンティティから連想されるエンティティを生成しており，多様性の向上は限定される．そこで本研究では，元のエンティティに捉われない多種多様なエンティティを生成し，より多様なエンティティの検知に対応することを目指す．具体的には，生成 AI を用いて元の文とは独立に架空のエンティティを生成し，元の文のエンティティと置き換える (図 2)．

5　実験

提案するデータ拡張手法によって作成したデータを用いて機密情報検知モデルを学習することにより，機密情報検知の精度が向上するかを確かめる．

5.1　実験設定

データセット　データセットにはストックマーク株式会社が公開する Wikipedia を用いた日本語の固有表現抽出データセット [18] を使用した．データセットの分割は huggingface の llm-book/ner-wikipedia-dataset[1] と同様のものを使用した．なお，本研究では「法人名」のラベルを検知対象の企業・法人名として扱う．そのため，各事例について「法人名」のラベル以外を全て「O」ラベルに変換する前処理を行った．

表 2　機密情報検知モデルの評価結果．

	適合率	再現率	F1	F2
ベースライン	0.899	0.863	0.881	0.870
MR-random [4]	0.886	0.879	0.883	0.880
MR-GPT [7]	0.887	0.883	0.885	0.884
周辺文脈の拡張	0.900	0.867	0.883	0.873
エンティティの拡張	**0.917**	**0.895**	**0.906**	**0.900**

データ拡張　4 節に示す各手法について，元のデータセットの学習データ，検証データに含まれる正例 1 件あたり 4 件の疑似データを作成した．実際に作成したデータの例を表 1 に示す．なお，学習データ量の増加により提案手法が不当に有利になることを防ぐため，ベースラインには元のデータセットのうち法人名ラベルが含まれる事例のみ 5 倍に複製したものを用いた．また，比較手法として Dai ら [4] の提案する MR (MR-random)，Sharma ら [7] の提案する GPT を用いた MR (MR-GPT)[2] も検証した．最終的に，学習データの事例数は 9642，検証データの事例数は 1166 となった．なお，評価データはデータ拡張を行わず，元の評価データと同量の 535 事例を用いた．

データ拡張に用いる生成 AI　データセットの拡張には Azure OpenAI の gpt-3.5-turbo-16k を使用した．生成されるデータの多様性を向上させるため，サンプリング温度は周辺文脈の拡張では 1.6，エンティティの拡張では 1.5 とした．また，MR-GPT では Sharma らと同様に 0.8 とした[3]．

1)　https://huggingface.co/datasets/llm-book/ner-wikipedia-dataset
2)　モデルには gpt-3.5-turbo-16k を用いた．プロンプトは Sharma らのプロンプトを日本語に翻訳したものを使用した．
3)　その他のパラメータはデフォルトのものを使用した．

表 3 評価データにおける予測事例．太字は検知対象の企業・法人名を示す．

文	ベースライン	周辺文脈の拡張	エンティティの拡張
日活テレビ映画芸術学院 として映画会社の **日活** が経営し、1975 年に日活撮影所内の一角に設立された。	日活、日活	日活テレビ映画芸術学院、日活	日活テレビ映画芸術学院、日活
その後に報じられた **ブルームバーグ** の記事では、この盗聴が 2000 年 6 月に既に始まっていた可能性があると伝えられた。	ブルームバーグ	なし	ブルームバーグ

機密情報検知モデル　機密情報検知のモデルには BERT-CRF を使用した．BERT-CRF は BERT [1] に CRF 層を組み合わせたモデルで，BERT 単独のモデルと比較して固有表現抽出の精度が高いことが知られている [19]．BERT には東北大学が公開する事前学習済みの日本語 BERT[4] を使用した．各データセットで 5 エポック学習し，評価には検証データでの F1 が最大のチェックポイントを用いた[5]．

5.2 評価方法

評価指標には適合率，再現率，F1，および F2 を用いた．機密情報を検知する上では検知漏れ，すなわち偽陰性は重大な問題である．そのため，再現率を重視する指標である F2 も評価指標として使用する．

5.3 実験結果

機密情報検知モデルの評価結果を表 2 に示す．表 2 から，エンティティの拡張が全ての指標で最も性能が良く，ベースラインと比較して F1 で 2.5%の大幅な精度向上を確認した．拡張元のデータセットには，例えば「トヨタ」や「マツダ」のような片仮名のみで構成されるエンティティが多数含まれており，表層的な特徴を丸暗記するようなショートカット学習 [20] ができてしまう可能性がある．エンティティの拡張を行ったモデルでは，多種多様な表現が生成されたことにより，表層的な特徴を用いた判断が難しくなり，より精緻な文脈理解が必要になったことで，精度向上につながったと考えられる．また，エンティティの拡張が MR-random，MR-GPT と比較して性能が良いことから，事実性に捉われない多様なエンティティを生成することは機密情報検知の精度向上に有効であると考えられる．さらに，周辺文脈の拡張についても，エンティティの拡張には及ばないものの，ベースラインと比較して全ての指標で性能が向上することを確認した．

4） https://huggingface.co/cl-tohoku/bert-base-japanese-v3

5） 学習率は 1e-4，バッチサイズは 32，warmup_ratio は 0.1 とした．

6 分析

本節では機密情報検知モデルの評価結果を定性的に分析する．表 3 に評価データにおける予測事例を示す．一つ目の事例では，提案するデータ拡張によって「日活テレビ映画芸術学院」という複雑な法人名を検知できていることが確認できる．この結果から，多様な文脈・エンティティを学習したことでモデルが文字種の複合を含む複雑な企業・法人名に対応できるようになったことが示唆される．一方，二つ目の事例では周辺文脈の拡張を行ったデータで学習したモデルが「ブルームバーグ」を検知できなかった．この事例をはじめ，表層からは人名との判別が難しい表現で検知に失敗する場合が存在した．データ拡張の際，地名や人名に似たエンティティでは，適切に企業・法人名の文脈で用いられていない例が散見された．そのような例を学習データから除去することで，周辺文脈の拡張においてもさらに精度を向上させることができると考える．

7 おわりに

本研究では，機密情報検知の精度向上に向けて，生成 AI を用いたデータ拡張を提案した．具体的には，エンティティ以外の周辺文脈を拡張する手法，エンティティ自体を拡張する手法を提案した．実際に拡張したデータを用いて機密情報検知モデルを学習した結果，エンティティを拡張する手法ではベースラインに比べ F1 で 2.5%の大幅な精度向上を確認した．本研究では企業・法人名に着目したが，今後は人名等の他の機密情報についても同様の手法の有効性を検証していく予定である．また，本研究で提案した生成 AI を用いた事実性に捉われないデータ拡張は機密情報検知以外の自然言語処理タスクにも応用できる可能性がある．したがって，その他のタスクに対する提案手法の有効性の調査も今後の課題としたい．最後に，今回の手法を改良して機密情報検知システムへの適用を進めていく予定である．

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 Proceedings of NAACL, pp. 4171–4186, 2019.

[2] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In Findings of ACL-IJCNLP, pp. 968–988, 2021.

[3] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. AI Open, pp. 71—90, 2022.

[4] Xiang Dai and Heike Adel. An analysis of simple data augmentation for named entity recognition. In Proceedings of COLING, pp. 3861–3867, 2020.

[5] 寺本優香, 駒水孝裕, 波多野賢治. 固有表現タグおよび pos タグによる交換制約付きデータ拡張手法. DEIM2023 最終論文集, 2023.

[6] Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In Proceedings of ACL, pp. 7056–7066, 2020.

[7] Saket Sharma, Aviral Joshi, Yiyun Zhao, Namrata Mukhija, Hanoz Bhathena, Prateek Singh, and Sashank Santhanam. When and how to paraphrase for named entity recognition? In Proceedings of ACL, pp. 7052–7087, 2023.

[8] Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. MELM: Data augmentation with masked entity language modeling for low-resource NER. In Proceedings of ACL, pp. 2251–2262, 2022.

[9] Claude Coulombe. Text data augmentation made simple by leveraging nlp cloud apis. In arXiv:1812.04718, 2018.

[10] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Proceedings of EMNLP-IJCNLP, pp. 6382–6388, 2019.

[11] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Proceedings of ACL, pp. 86–96, 2016.

[12] Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In Proceedings of EMNLP, pp. 1268–1283, 2020.

[13] Jian Liu, Yufeng Chen, and Jinan Xu. Low-resource ner by data augmentation with prompting. In Proceedings of IJCAI, pp. 4252–4258, 2022.

[14] Zhu Wenjing, Liu Jian, Xu Jinan, Chen Yufeng, and Zhang Yujie. Improving low-resource named entity recognition via label-aware data augmentation and curriculum denoising. In Proceedings of CCL, pp. 1131–1142, 2021.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Proceedings of NeurIPS, pp. 3104–3112, 2014.

[16] 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 Proceedings of NeurIPS, pp. 1877–1901, 2020.

[17] Satoshi Sekine and Hitoshi Isahara. IREX: IR & IE evaluation project in Japanese. In Proceedings of LREC, 2000.

[18] 近江崇宏. Wikipedia を用いた日本語の固有表現抽出のデータセットの構築. 言語処理学会年次大会発表論文集, pp. 350–352, 2021.

[19] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using BERT-CRF. In arXiv:1909.10649, 2019.

[20] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. Nature Machine Intelligence, pp. 665—673, 2020.

A　データ拡張で用いるプロンプトの詳細

本節では，提案するデータ拡張手法において実際に生成 AI に入力したプロンプトの詳細について述べる．

A.1　周辺文脈の拡張

周辺文脈の拡張で用いたプロンプトを表 4 に示す[6]．文生成のプロンプトの *LOCATION* には，「できる限り文末」と「文の先頭」を 9 対 1 の比率で使用し，文におけるエンティティ位置の多様性向上を目指した．

また，低品質な学習データの生成を抑制するため，生成結果を再度生成 AI に入力し校正を行った．具体的には表 4 に示すプロンプトを用いて，文法誤りが存在する場合にはそれを訂正した文，非文の場合には「×」を出力させた．なお，文校正の際は，生成 AI のサンプリング温度は 0.1 とした．データ作成時には，各事例について，非文と判定されない文が 4 つ生成されるまで生成を繰り返した[7]．

A.2　エンティティの拡張

エンティティの拡張で用いたプロンプトを表 5 に示す．多様なエンティティを生成するために，日本語の企業・法人名を生成するためのプロンプトと英語の企業・法人名を生成するためのプロンプトを 9 対 1 の比率で併用した[8]．日本語の企業・法人名を生成するためのプロンプトには「漢字・カタカナ・ひらがなを自由に使用してください。」と記載することで，日本語の中でも様々な文字種を使用して生成させることを促している．また，英語の企業・法人名については，生成結果を 1 対 1 対 1 の比率で「そのまま使用」「文字列を大文字に変換」「文字列を小文字に変換」する処理を行い，さらなる多様性向上を目指した．

表 4　周辺文脈の拡張で使用したシステムプロンプト．*ENTITY* は文に含まれるエンティティを示す．

文生成	非文判定・文法誤り訂正
次に示す企業/法人名を含む一文を生成してください。 企業/法人名は *LOCATION* で生成してください。 文に含まれる企業/法人名は与えられた企業/法人名のみとしてください。 企業/法人名は以下の形式で与えられます。 ‘‘‘ {”entity”: ”企業/法人名”} ‘‘‘	次の文が自然な日本語となるように修正してください。 ### 修正のルール ・修正は「て」「に」「を」「は」の間違いなど軽微なものに限る ・入力文が日本語ではない場合は、”×” のみを出力する ・以下の企業/法人名は書き換えない ### 企業/法人名 ・*ENTITY*

表 5　エンティティの拡張で使用したユーザプロンプト．

エンティティの拡張 (日)	エンティティの拡張 (英)
企業/法人名を一つ生成してください。 以下の例のように漢字・カタカナ・ひらがなを自由に使用してください。 ### 企業/法人名の例 ‘‘‘ ワンダー自動車 たけぼうグループ 富士山電機 アルファ・キャピタル 未来通信 ‘‘‘ ### 企業/法人名の例 終了	Please give me an idea about the name of a new company (or an organization). Provide only ”one” name. Feel free to use lowercased / uppercased letters as shown in the example below ### Example ‘‘‘ EPIC Zevo Inc. ETA VitaNexus Elevate Industries ‘‘‘ ### End of example

6)　ユーザプロンプトについては文生成では企業・法人名を，非文判定・文法誤り訂正では文生成で生成した文を与える．なお，複数の企業・法人名を含む文を生成する際には企業・法人名を箇条書きで与え，プロンプトの一文目を「次に示す企業/法人名を順に使って一文を生成してください。」に変更した．

7)　繰り返し回数が 20 回を超えた事例は生成 AI によるデータ作成が難しいと判断し，データセットから除去した．

8)　20 文字を超えるエンティティは企業・法人名として不自然であると判断し除外した．