攻撃者の振る舞い抽出のための遠距離教師あり学習

山嵜 麿与[1,a)]

概要：効果的なインシデント対応のために脅威レポートの共有が行われているが，自然文で記述された膨大な知識を人手で利活用することは容易ではない．このため，教師あり学習を用いて自然文から攻撃者の振る舞い情報を抽出する手法が検討されている．しかし，人手による教師データの作成コストが高いために，利用可能な教師データが不足しており，高精度な抽出器が実現できていない．そこで本研究では，攻撃者の振る舞いに関連する攻撃手法の名称や観測事象等の特徴語に着目し，人手によるラベル付け無しに大規模な擬似教師データを作成可能な遠距離教師あり学習手法を提案する．文に対する攻撃者の振る舞いのマルチラベル分類を行う評価実験の結果，擬似教師データとノイズモデリングネットワーク用いることにより，従来手法の F1 値を 0.29 上回る精度の 0.82 で攻撃者の振る舞いを抽出可能であることを示す．また，大規模な脅威レポートから抽出した攻撃者の振る舞いの共起関係の可視化により，提案手法を用いることで，インシデント対応等の業務支援に活用可能な知見が得られることを示す．

キーワード：脅威インテリジェンス，攻撃者の振る舞い抽出，遠隔教師あり学習

Distantly Supervised Learning for Adversary Behavior Extraction

Mayo YAMASAKI[1,a)]

Abstract: Although threat reports are shared for effective incident responses, it is difficult to utilize vast knowledge of natural language reports manually. Therefore, it has been proposed that supervised machine learning-based extraction methods for adversary behaviors. However, these methods suffer from low performance because of a shortage of labeled data with high costs to develop manually. This paper proposes a distantly-supervised learning method that utilizes attack method names and observables related to adversary behaviors and creates large pseudo labeled datasets without human annotations. On a multi-label sentence classification task, this paper experimentally shows that the proposed method with the pseudo labeled data and noise modeling networks achieved an F1-score of 0.82, which is 0.29 higher than that of the conventional method. Furthermore, visualization of co-occurrence relationships in adversary behaviors extracted from large scale threat reports shows that the proposed method should be useful to obtain intelligence to be used to support operations such as incident response.

Keywords: Threat Intteligence, Adversary Behavior Extraction, Distant Supervision

1. はじめに

攻撃者の振る舞い抽出タスクは，予め定義された攻撃者の振る舞いの内，与えられた文書で言及されている振る舞い集合を抽出するタスクである．攻撃者の振る舞いとしては，ATT&CK Technique[*1]が用いられ，マルチラベルの文書分類問題として扱われる [1], [2]．図 1 の例は，一文からなる文書が与えられた際に出力される攻撃者の振る舞い集合である．

現在の脅威インテリジェンス共有では，構造化が容易な IoC(Indicator of Compromise) 情報は MISP[*2]や

[1] NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories
[a)] mayo.yamasaki.ua@hco.ntt.co.jp

[*1] https://attack.mitre.org/
[*2] https://www.misp-project.org/

Input: Hackers sent emails containing URLs linked to Word files with malicious macro that execute PowerShell script.

Output: Spearphishing Link, User Execution, PowerShell, Scripting

図 1　攻撃者の振る舞い抽出タスク
Fig. 1　Adversary Behavior Extraction Task

STIX/TAXII[*3]の様な構造化形式を用いて共有されている一方で，構造化が困難な攻撃者の振る舞いに関する情報はレポートや電子メール等の自然文形式で共有されている [3]．このため，自然文で記述された攻撃者の振る舞いに関する知見への可触性が低く，脅威インテリジェンスを十分に利活用できていない問題がある．この問題を解決するため，教師あり学習を用いて自然文から攻撃者の振る舞い情報を抽出する手法が検討されているが，学習に利用可能な教師データが不足しており，高い精度での抽出が実現できていない．

そこで本研究では，教師データの不足に対処するために，関係抽出タスクで提案された遠距離教師あり学習（Distantly Supervised Learning）[4] を応用することで，人手によるラベル付け無しに大規模な擬似教師データを作成する手法を提案する．遠距離教師あり学習は，ラベルの無い文書に対して知識ベースやルールを用いてノイズを含む擬似教師を付与し学習に用いる手法で，関係抽出 [4] や感情分析 [5] タスクにて用いられているが，これまで，攻撃者の振る舞い抽出タスクにおいては検討されていない．提案手法では，擬似教師を付与する対象を脅威レポートに限定し，各攻撃者の振る舞に関連する攻撃手法の名称と、それらを実行する上で用いられるコマンド名，プログラム名，API 名等の観測事象に関連する特徴語を用いることで，高い精度での擬似教師の付与を可能にする．本研究の貢献は以下の通りである．

- 攻撃者の振る舞い抽出タスクに遠隔教師あり学習手法を適応した初めての手法を提案する．評価実験により，単純なルールを用いることで高い精度で疑似教師を付与可能であること，また，付与された疑似教師データを用いることで，既存手法を上回る精度での攻撃者の振る舞い抽出が可能であることを示す．
- 疑似教師データには偽陽性・偽陰性のノイズが存在しているため，画像処理分野で提案されているノイズモデリングネットワーク [6] を，文書分類に応用する手法を新たに提案する．ノイズモデリングネットワークは，クリーンデータ無しにノイズによる影響を軽減する手法で，提案手法の適応により，本タスクでの抽出精度が向上することを実験により示す．
- 大規模な脅威レポートから「initial-access」と「execution」に関する攻撃者の振る舞いを抽出し，振る舞い間の共起関係に関する可視化を行った．可視化により，共起関係が専門家の経験と一致することを示し，また，提案手法によりインシデント対応等の業務に活用可能な知見が得られることを示す．

[*3] https://oasis-open.github.io/cti-documentation/

2. 提案手法

遠隔教師あり学習を用いた攻撃者の振る舞い抽出器は，脅威レポートの収集，疑似教師の付与，モデルの学習の 3 つの段階で構築される．本抽出器では，疑似教師のノイズを削減するため，また，より詳細な情報抽出を行うために，文単位での教師の付与と分類モデルの学習を行った．

2.1 脅威レポートの収集

疑似教師の付与対象として利用可能な脅威レポートのデータセットが存在していないため，クローラを用いて新たにデータセットを作成した．収集対象は脅威レポートを公開しているセキュリティベンダ・メディアの 34 の WEB サイトで，scrapy[*4]を用い，robot.txt の内容に従い 2020 年 4 月 27 日時点で取得可能な全てのページを収集した．ただし，評価実験におけるリーケージを避けるため，ATT&CK から参照されているすべての URL を除外した．また，各サイトの全てのページが脅威レポートではないため，各サイト毎のカテゴリ，タグ情報を基に，脅威レポートであるか否かのフィルタリングを実施した．その後，python-readability[*5]を用いて HTML から本文を取得し，spacy[*6]による文分割を実施した．

クロール，本文抽出，文分割のエラーによって生じた意味をなさない文を除外するために，文字長を用いたフィルタリングを実施した．文字長の下限を 25 文字，上限を四分位範囲の 2 倍に相当する 288 文字とし，この範囲内の文のみを利用した．これらのフィルタリングの結果，19,259 の脅威レポートに紐づく 667,431 文を収集した．

2.2 疑似教師の付与

ATT&CK は攻撃者の振る舞いに関する知識ベースであり，過去の実例から ATT&CK Technique として，攻撃者の振る舞いが定義されている．各振る舞いには，具体例を含む説明文と，過去の攻撃事例でどのように利用されたかの要約等が付与されている．疑似教師の作成には抽出規則を人手で作成する必要があり，教師データの付与コストと比較して少ないコストではあるものの時間を要する．このため本研究では，比較的重要度が高いと考えられる「initial-access」と「execution」に紐づく ATT&CK Technique のみを抽出対象とした．これらの振る舞いに関する疑似教師を付与するために，攻撃手法の名称と、それらを実行する上で用いられるコマンド名，プログラム名，

[*4] https://scrapy.org
[*5] https://github.com/buriy/python-readability
[*6] https://spacy.io/

API 名等の観測事象に関連する特徴語に着目し，(1) フレーズ規則，(2) 正規表現規則，(3) CPE[*7]規則，(4) 複合規則の 4 つの規則を用いた．(1) ～ (3) への適合結果は振る舞いと直接対応しておらず，それらを組み合わる複合規則により振る舞いに対応付けられるものも存在する．

フレーズ規則では，特徴的なフレーズとトークン化後のトークン列が一致した場合に，特徴語と関連する攻撃者の振る舞いがその文で言及されていると仮定する．フレーズには，攻撃手法の名称 (e.g. drive　by, spearphishing)，コマンドの名称 (e.g. cron, WMIC)，プログラムの名称 (e.g. Regsvcs.exe, schtasks.exe)，API の名称 (e.g. LoadLibrary, CreateProcessA)，その他の特徴語 (e.g. compiled html, rdp, facebook, chrome) を用いた．

正規表現規則は，フレーズ規則で抽出困難な場合に用いた．例えば，プログラム名や API の名称には，特徴語の前後に任意の文字列が連続する可能性があるため，「(^|\s)(.*[\\/ ¥])?Regsvcs(\.exe)?(\s |$)」や「(^|\s)CreateProcess(.*)?(\s|$)」等の正規表現を用いた．

CPE 規則では，文中の CVE が影響する CPE に基づき，文中で言及されている製品種別の抽出を行った．抽出対象に含まれる脆弱性侵害に関する振る舞いである「Exploit Public-Facing Application」と「Exploitation for Client Execution」を識別するために，製品種別として client application (e.g. flash player, firefox) と public service (e.g. nginx, drupal) を用いた．5,955 文で CVE に対する言及があり，言及頻度の上位 163 件の CPE (CVE への言及の 89.25%) のみを付与対象とした．

複合規則では，これらを組み合わせた抽出を行う．例えば，「spearphishing」は特徴語であるものの ATT&CK で定義される「Spearphishing Attachment」,「Spearphishing Link」,「Spearphishing via Service」のいずれに該当するかを判別することはできない[*8]．このため，同一文内に「link」や「attachment」への言及が存在するか否かをフレーズ規則により判定し，抽出を行う．例えば，「spearphishing」を含みかつ「attachment」を含む文に対しては，「Spearphishing Attachment」を付与する．

ATT&CK の内容は日々更新されており，ルールの作成時点で，「initial-access」と「execution」に含まれる ATT&CK Technique は 45 個存在していた．ただし特徴語が無く，構文や意味情報を用いない単純な規則では抽出困難であった 6 個 (Trap, Space after Filename, Hardware Addition, Trusted Relationship, Source, Signed Script Proxy Execution) の振る舞いについては，抽出対象から除外した．従って，39 個の振る舞いに対して抽出規則を作成し，収集した全ての文に対して抽出規則を適応した．適応結果を図 2 に示す．疑似教師が付与された文は 74,174 文 (付与候補の 11.1%) であり，付与された文の内ラベル数 1 個の文は 58,782 文，2 個は 13,748 文，3 個は 1,496 文，4 個は 133 文，5 個は 14 文，6 個は 1 文であった．

図 2　攻撃者の振る舞い毎の文数
Fig. 2　Number of Sentences each Adversary Behavior

2.3　モデルの学習

単純な規則に基づき付与した疑似教師データにはノイズが存在する．しかし，人手によるアノテーションを行った場合においてもノイズは存在し，特に，クラウドソーシングを用いて不特定多数がアノテーションを行った場合，アノテーションの一貫性を保つことは困難である．このため，遠隔教師あり学習に限らずノイズを考慮したモデル化が検討されている．遠隔教師あり学習を用いた攻撃者の振る舞い抽出の場合，疑似教師に偽陽性と偽陰性が存在し，ノイズ分布の推定に利用可能なクリーンデータが存在しない．そこで，この条件下でマルチラベル分類を行うために提案されたノイズモデリングネットワーク [6] を文分類タスクに応用する．提案手法におけるモデルを図 3 に示す．

Φ を入力となる特徴量の集合，$z^c \in \{0,1\}$ をラベル c のノイズを含む観測値，$y^c \in \{0,1\}$ を観測できない真のラベル c の値とする．ノイズモデリングネットワークでは，擬

*7　https://nvd.nist.gov/products/cpe
*8　ATT&CK Sub Techniques 導入後，これらは「Phishing」という新たな technique の sub technique として定義されている．このため「Phishing」であることまでは判別可能である．

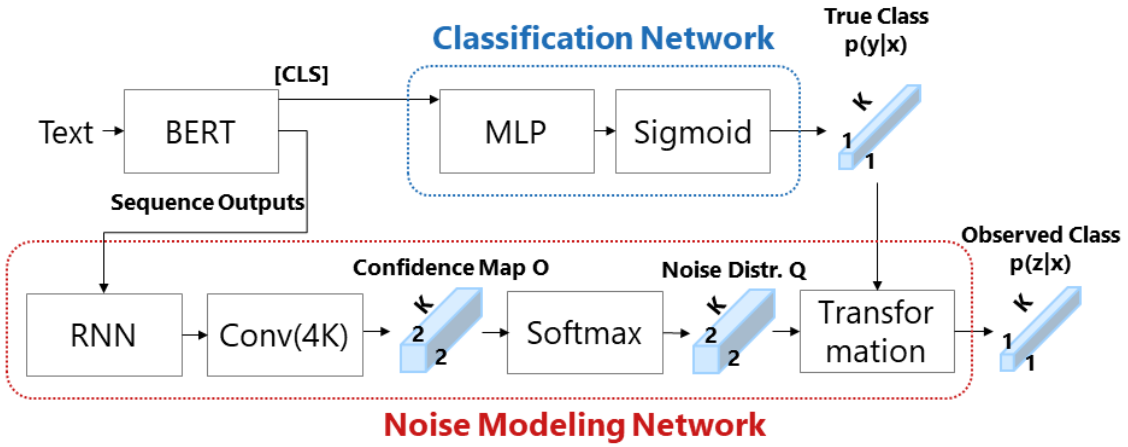

図 3　提案モデル
Fig. 3　Proposed Model

似教師データに存在する偽陽性と偽陰性のノイズを，ノイズ分布 $Q = p(z^c = i|y^c = j, \Phi)$ をモデル化することで対処する．図 3 に示す通り，提案モデルは，$p(y^c|\Phi)$ をモデル化する識別ネットワークと，$p(z^c|y^c, \Phi)$ をモデル化するノイズモデリングネットワークから構成される．これらのネットワークは 2 段階で学習される．1 段階目でノイズを含む観測ラベルを用いて識別ネットワークを学習し，2 段階目ではノイズモデリングネットワークを追加し，ネットワーク全体のパラメータを最適化する Fine-tuning を行う．

$h(\Phi)$ を入力の特徴ベクトルを出力する GRU(Gated Recurrent Unit)[7] による非線形変換とし，u_{ij}^c と b_{ij}^c をパラメータとした場合，ノイズ分布は次の式で定義される．

$$p(z^c = i|y^c = j, \Phi) = \frac{\exp(o_{ij}^c)}{\sum_i \exp(o_{ij}^c)} \tag{1}$$

$$o_{ij}^c = (u_{ij}^c)^T h(\Phi) + b_{ij}^c \tag{2}$$

o_{ij}^c は，真のラベル j から観測ラベル i への遷移スコアで，図 3 では確信度マップ O に相当する．式（1）を用いることで，観測ラベルの予測確率は次式で定義される．

$$p(z^c|\Phi) = \sum_j p(z^c = i|y^c = j, \Phi) p(y^c|\Phi) \tag{3}$$

モデルパラメータは，$p(z^c|\Phi)$ と観測ラベルの交差エントロピーを最小化することにより求める．即ち目的関数 L は次式で表される．

$$L = \sum_c z^c \log p(z^c|\Phi) + (1 - z^c) \log(1 - p(z^c|\Phi)) \tag{4}$$

擬似教師データは，図 2 に示す通り不均衡データであるため，事例数に応じた重み付けを行う．ラベル c の値 j に対する重み w_j^c は，訓練事例数 n，$z^c = j$ を満たす事例数 n_j^c とした場合に，次式で表される．

$$w_j^c = \frac{n}{2n_j^c} \tag{5}$$

式 (4),(5) より，最終的な目的関数 L は次式で表される．

$$L = \sum_c w_{z^c}^c \{z^c \log p(z^c|\Phi) + (1 - z^c) \log(1 - p(z^c|\Phi))\} \tag{6}$$

識別ネットワークは，BERT[8] の [CLS] トークンに対応する隠れ層に多層パーセプトロン (MLP) 層を結合したネットワークで，Fine-tuning を行い学習する．ノイズモデリングネットワークでは，BERT の各入力トークンに対する隠れ層に GRU を結合したネットワークで特徴ベクトル $h(\Phi)$ を計算する．この特徴ベクトルは o_{ij}^c で定義される様にラベル毎に 4 つの線形変換が行われる，ラベルの種類を K とした場合，サイズ 1×1 の $4K$ 個のカーネルを持つ畳み込み層 $Conv(4K)$ として表現できる．畳み込み層の出力である確信度マップ O は $\mathbb{R}^{2\times2\times K}$ に変形され，softmax 層を通じてノイズ分布 Q が算出される．$p(z^c|\Phi)$ は，式 (3) に示す通り，$p(y^c = 0|\Phi)$ と $p(y^c = 1|\Phi)$ のノイズ分布 Q による重みづけ和を計算することで算出される．

ネットワーク全体の最適化後は，ノイズモデリングネットワークを用いずに識別ネットワークのみを用いて真のラベル y^c の予測を行う．

3. 実験

遠隔教師あり学習を用いた攻撃者の振る舞い抽出の精度を評価するための実験を行った．

3.1 データセット

学習データとして用いる疑似教師の付与を行った文の内，事例数が 10 以下の 3 個のラベル（Windows Remote Management, AppleScript, CMSTP) については，訓練と評価の対象から除外した．また学習データに負例が存在

しないため，疑似教師の付与規則に適合しなかった文から 2343 文 (1 ラベル当たりの平均事例数) を無作為に抽出し学習に用いた．学習データの 80%(61,213 文) を訓練データに，残りの 20%(15,304 文) を検証データとして用いた．ラベル毎の事例数を維持したまま分割するために，scikit-multilearn[*9]を用いて分割を行った．

利用可能な評価データが存在しないため，2 種類のデータセットを新たに構築し評価実験を行った．これらのデータセットは文書に対して振る舞いが付与されているため，文書単位での抽出精度の評価を実施した．

CAED（Complemented ATT&CK Examples Dataset）: ATT&CK Technique には数文からなる攻撃事例が複数記述されている．事例は教師有り学習に用いることを意図していないため，当該 technique 以外にも該当する可能性がある．CAED は，technique 毎に最大 20 件までの事例を無作為抽出し，当該 technique 以外で言及されている technique を付与したデータセットである．合計 388 文書（平均 1.1 文）からなり，新たに 238 のラベルを付与している．CAED は，攻撃事例が存在した 33 の振る舞いについての Precision, Recall, F 値での精度評価が可能であるが，ATT&CK の語彙に合わせて記述されているため比較的抽出が容易であることに注意されたい．

FARD（Filtered ATT&CK Reference Dataset）: ATT&CK Technique には外部の脅威レポート等のリファレンス URL の一覧が付与されている．FARD は，各 URL の本文に対して，同一 URL を参照している technique を付与したデータセットである．PdfAct[*10]を用いて PDF から本文抽出を行い，HTML からの本文抽出と本文の文分割には疑似教師付与と同じ手順を適応した．本文抽出と文分割のエラーが存在していたため，意味のある文 (50 文字以上) を 10 文以上含む文書のみを用いた．また，明らかに振る舞いの付与が不足している URL を除外するために，5 つ以上振る舞いが付与されていた 336 文書（平均 324.1 文）のみを評価に用いた．FARD は，リファレンスを含む 32 の振る舞いについて実際の脅威レポートを用いた評価が可能であるが，当該文書の全ての振る舞いが付与されているわけではないため Recall のみを評価可能である．

3.2 実験条件

提案モデルの実装には，Keras[*11]と huggingface transformers[*12]を用い，BERT の事前学習済みモデルには bert-base-uncased を用いた．また，提案モデルの最適化には Adam を $\beta_1 = 0.9$，$\beta_2 = 0.999$ で，STLR[9] を $lr\_max = 2e-5$, $cut\_frac = 0.1$, $ratio = 32$ で用いた．その他のハイパーパラメタとして，バッチサイズを 32，MLP 層の隠れ層を 300，dropout rate を 0.1，入力の最大長を 128 とした．

評価実験には以下の 4 つのモデルを用いた．

- **rcATT:** 特徴量に BoW の tf-idf を用いた Linear SVM モデルで，ATT&CK のリファレンスの 1490 文書を用いて学習された実装[*13]を用いた．このため、リーケージを避けるために，CAED での評価のみに用いた．
- **DS Rule:** 疑似教師の付与規則のみを用いた手法．
- **BERT+MLP:** 提案モデルの識別ネットワークのみを用いたモデル．epoch 数を 15 とした．
- **BERT+MLP+NMN:** 図 3 の提案モデル．識別ネットワークの学習 epoch 数を 10，ノイズモデリングネットワークを含むネットワーク全体の学習 epoch 数を 5 とした．

3.3 実験結果

表 1 に CAED，表 2 に FARD の実験結果を示す．いずれの結果においても，提案手法である BERT+MLP と BERT+MLP+NMN が、既存手法の rcATT を大きく上回る精度を達成している．これは，利用可能な教師データの量とモデルアーキテクチャによる差異を表している．

ノイズモデリングネットワークの導入より，CAED での評価では Precision と F 値が向上しているものの Recall の減少が見られ，FARD での評価においても Recall の減少が見られた．従って，ノイズのモデル化により, 疑似教師の偽陽性の影響が軽減されていると見られる．

CAED での評価では，DS Rule が機械学習を用いた手法の Precision を大きく上回っている．これは，CAED の文書が ATT&CK の語彙を用いて記述されているため、付与規則を用いた抽出が容易であるためである．しかし規則化できない事例も存在するため，Recall については機械学習手法を下回る結果となっている．

表 1 と表 2 を比較すると，Recall の micro 平均の差異は少ないものの，macro 平均では FARD で減少がみられる．これは，一部のラベル分類においてモデルの汎化性能が低く，実レポートにおける多様な表現からの抽出が困難であることを示唆している．

4. 攻撃者の振る舞いの共起関係

大規模な脅威レポートに対する攻撃者の振る舞いの共起関係に関する調査を実施した．調査では提案手法を用いて，2.1 で収集した 667,431 の文から攻撃者の振る舞い抽出を行い，15,656 文書に紐づく 111,556 文を抽出した．同一文書に含まれる振る舞いのすべてが同一の攻撃で観測されたものと仮定し，共起する振る舞いの分析を実施した．図 4 は，

[*9] http://scikit.ml/index.html
[*10] https://github.com/ad-freiburg/pdfact
[*11] https://keras.io
[*12] https://github.com/huggingface/transformers
[*13] https://github.com/vlegoy/rcATT

表 1　CAED での評価結果
Table 1　Evaluation Results on CAED

	Precision		Recall		F 値	
	Micro	Macro	Micro	Macro	Micro	Macro
rcATT	0.53	0.50	0.56	0.63	0.55	0.53
DS Rule	**0.98**	**0.97**	0.78	0.77	0.87	**0.84**
BERT+MLP	0.88	0.82	**0.86**	**0.83**	0.87	0.80
BERT+MLP+NMN	0.91	0.86	0.85	**0.83**	**0.88**	0.82

表 2　FARD での評価結果
Table 2　Evaluation Results on FARD

	Recall	
	Micro	Macro
DS Rule	0.79	0.67
BERT+MLP	**0.87**	**0.73**
BERT+MLP+NMN	0.85	0.72

行の振る舞いが観測された際に列の振る舞いが観測される確率を可視化したもので，各振る舞いは，「initial-access」と「execution」の順に分けて整列されている．また，濃い色は確率が高く，薄い色は確率が低いことを示している．

図 4 より，「External Remote Services」が観測された際に「Valid Accounts」が共起していることや，「User Execution」が観測された際に「Spearphishing」が共起していることがわかる．また，濃い色の列はいずれの振る舞いとも共起している振る舞いを意味しており，例えば「initial-access」としては，「Valid Account」や「Spearphishing Attachment」が頻繁に組み合わされる振る舞いであることがわかる．このように，図 4 の共起関係は専門家の経験と一致することを示唆している．従って，より多くの攻撃者の振る舞いに対する共起関係を分析を行うことで，専門知識が不足している運用者のインシデント対応等の業務支援システムの構築に有益であると考えられる．

5. 関連研究

5.1 攻撃者の振る舞いの抽出

脅威レポートを対象とした情報抽出では，IoC[10], [11], [12]，マルウェア [13], [14]，攻撃キャンペーン全体 [15], [16] を対象としたものがある．情報抽出では，抽出対象の情報種別に適合する文書の部分列を抽出するため，未知の情報を取得できる利点があるものの，抽出した情報の正規化が必要になる．一方で，ATT&CK で定義される攻撃者の振る舞いの様に，予め抽出対象の情報を限定することで，未知の振る舞いの抽出は困難になるが正規化が不要となる．ATT&CK を対象とした攻撃者の振る舞いの抽出では，ルールベースの手法 [1] と教師あり学習を用いた手法 [2] があり，いずれも文書分類の問題として解決している．

教師データは ATT&CK のリファレンス情報を基に作成されているが，教師あり学習のためのデータを想定したものではないため，ラベルが不完全で多数の偽陰性のノイズを含む教師データとなっている．

5.2 事前学習モデル

近年の NLP では，目的タスクでの学習データの不足に対処するため，予め一般的な言語表現を大規模なデータで学習し，獲得した言語表現を目的のタスクに転移する手法が広く用いられている．深層学習を用いて事前学習した表現を特徴量として用いる手法 [17], [18] や，文脈に基づいた表現を事前学習し目的タスクに Fine-tuning する手法 [8], [19] がある．特に [8] の BERT は広く利用されており，より発展させた研究 [20], [21] や，文書分類タスク等への活用が行われている [22], [23].

5.3 遠距離教師あり学習

文書から実態間の関係を抽出する関係抽出タスクでは，教師データの不足に対処するために，遠距離教師あり学習と呼ばれる手法が提案されている [4]．遠距離教師あり学習は，知識ベースに存在する事実 (実態間の関係) を用いて，ある一文内に 2 つの実態に対する言及が存在し，かつ，知識ベースにそれらの実態間に関係が存在する場合に，その関係ラベルを擬似的な教師として付与するヒューリスティクスに基づいて学習を行う手法である．擬似的な教師には偽陽性と偽陰性が存在するものの，人手によるアノーテーション無しに，大規模な教師データの作成が可能になる．遠距離教師あり学習は，固有表現認識 [24] や感情分析 [5] においても異なるヒューステイクスによって作成した擬似的な教師データを用いた学習手法が検討されている．

5.4 ノイズ削減手法

遠距離教師あり学習では，擬似的な教師に存在するノイズ削減を行う手法が提案されている [25]．深層学習を用いた関係抽出タスクでは，2 つの実態を含む全ての文に対して関係への言及を仮定する仮説を緩和させ，偽陽性に対処する手法 [26] が提案されており，マルチラベルタスクにも適用されている [27]．偽陽性と偽陰性に対処する手法としては，ノイズモデル層を用いた文書分類が提案されているが [28]，マルチラベルタスクを対象としていない．

画像処理の分野でもノイズ削減の手法 [29] が提案されて

図 4　攻撃者の振る舞いの条件付き確率

Fig. 4　Conditional Probabilities of Adversary Behaviors

おり，[28] もこの手法を応用している．[6] は，[29] を発展させたもので，マルチラベルタスクでの偽陽性と偽陰性によるノイズを削減するためのノイズモデルネットワークを提案している．本研究では [6] の手法を元に，文書分類タスクでのノイズモデルネットワークを提案する．

6. おわりに

本稿では，遠隔教師あり学習を攻撃者の振る舞い抽出に用いた新たな手法を提案した．評価実験の結果，単純なルールによって付与された疑似教師データを用いることで，既存手法を大きく上回る精度で攻撃者の振る舞いを抽出可能であることを示した．また，提案手法によりセキュリティ運用にて利用可能な知見が得られることを示した．

参考文献

[1] Ghaith Husari, Ehab Al-Shaer, Mohiuddin Ahmed, Bill Chu, and Xi Niu. Ttpdrill: Automatic and accurate extraction of threat actions from unstructured text of cti sources. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pp. 103–115, 2017.

[2] Valentine Legoy, Marco Caselli, Christin Seifert, and

Andreas Peter. Automated retrieval of att&ck tactics and techniques for cyber threat reports. *arXiv preprint arXiv:2004.14322*, 2020.

[3] ENISA. Exploring the opportunities and limitations of current threat intelligence platforms, 2018.

[4] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics, 2009.

[5] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, Vol. 1, No. 12, p. 2009, 2009.

[6] Zhuolin Jiang, Jan Silovsky, Man-Hung Siu, William Hartmann, Herbert Gish, and Sancar Adali. Learning from noisy labels with noise modeling network. *arXiv preprint arXiv:2005.00596*, 2020.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[10] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhou Li, Luyi Xing, and Raheem Beyah. Acing the ioc game: Toward automatic discovery and analysis of open-source cyber threat intelligence. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 755–766, 2016.

[11] Ziyun Zhu and Tudor Dumitras. Chainsmith: Automatically learning the semantics of malicious campaigns by mining threat intelligence reports. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 458–472. IEEE, 2018.

[12] Shengping Zhou, Zi Long, Lianzhi Tan, and Hao Guo. Automatic identification of indicators of compromise using neural-based sequence labelling. *arXiv preprint arXiv:1810.10156*, 2018.

[13] Ziyun Zhu and Tudor Dumitraş. Featuresmith: Automatically engineering features for malware detection by mining the security literature. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 767–778, 2016.

[14] Swee Kiat Lim, Aldrian Obaja Muis, Wei Lu, and Chen Hui Ong. Malwaretextdb: A database for annotated malware articles. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1557–1567, 2017.

[15] Peter Phandi, Amila Silva, and Wei Lu. Semeval-2018 task 8: Semantic extraction from cybersecurity reports using natural language processing (securenlp). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 697–706, 2018.

[16] Roshni R Ramnani, Karthik Shivaram, and Shubhashis Sengupta. Semi-automated information extraction from unstructured threat advisories. In *Proceedings of the 10th Innovations in Software Engineering Conference*, pp. 181–187, 2017.

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[18] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[19] Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. *Technical report, OpenAI*, 2018.

[20] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5754–5764, 2019.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[22] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Docbert: Bert for document classification. *arXiv preprint arXiv:1904.08398*, 2019.

[23] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pp. 194–206. Springer, 2019.

[24] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 1524–1534. Association for Computational Linguistics, 2011.

[25] Alisa Smirnova and Philippe Cudré-Mauroux. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, Vol. 51, No. 5, pp. 1–35, 2018.

[26] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 148–163. Springer, 2010.

[27] Xiaotian Jiang, Quan Wang, Peng Li, and Bin Wang. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1471–1480, 2016.

[28] Ishan Jindal, Daniel Pressel, Brian Lester, and Matthew Nokleby. An effective label noise model for dnn text classification. *arXiv preprint arXiv:1903.07507*, 2019.

[29] Sainbayar Sukhbaatar, Joan Bruna, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. Training convolutional networks with noisy labels. *arXiv preprint arXiv:1406.2080*, 2014.