設備保全レポートにおける 大規模マルチタスク情報抽出データセットの構築

谷口　元樹　　高橋　拓誠　　谷口　友紀　　大熊　智子
富士フイルム株式会社
motoki.taniguchi@fujifilm.com

概要

情報抽出の既存のデータセットの課題は、その規模が小さいこと、アノテーションされているサブタスクの数が限られていること、対象テキストがフォーマルであるため抽出が比較的容易であることである。本論文では、これらの課題を解決し、より実応用に近いテストベッドとして工場設備の保全レポートを対象にした情報抽出データセットを構築する。我々のデータセットの特徴は、インフォーマルな書かれ方が比較的多い保全レポートを対象に、文分類・固有表現認識・モダリティ推定・関係抽出の４つの情報抽出サブタスクのラベルが約 1500 文書に付与されていることである。

1　はじめに

情報抽出タスクの目的はテキストから固有表現や固有表現間の関係などの情報を抽出し、構造化することである。構造化された情報は情報検索やデータ分析などの下流のアプリケーションに有用である。

情報抽出タスクは固有表現認識、関係抽出、事実性判定などの複数のサブタスクから構成されている。固有表現認識では ACE (Automatic Content Extraction) などの評価型ワークショップが古くから開催されているように、情報抽出は多くの研究がなされてきた。近年にはディープラーニングを用いた手法の隆盛に加えて、複数のサブタスクを同時に学習・推論する手法も盛んに取り組まれている。

情報抽出の既存データセットにはいくつかの問題がある。１つ目は、データセットの規模が小さいことである。ディープラーニングを用いた手法では、より高精度に情報を抽出できるモデルを学習するためには、データセットのサイズが大きいことが望ましい。２つ目は、多くのデータセットが１つもしくは２つの情報抽出のサブタスクのラベルしか付与されていないことである。抽出した情報を情報検索やデータ分析などの下流のアプリケーションで活用する実応用を考えた際には、一つのサブタスクで定義できる情報だけでは一面的な情報しか抽出できず、十分ではない。また、ディープラーニングを用いた手法では複数のタスクを同時に学習するマルチタスク学習を行うことで、より高精度なモデルを学習することが知られている。このため、多くのサブタスクを同一文書にアノテーションしたデータセットがあれば、より効率的なモデル開発を行うことができる。３つ目は、対象となるテキストの書かれ方である。既存の多くのデータセットにはニュース記事などのフォーマルなテキストが用いられている。フォーマルなテキストを対象にした情報抽出は高い精度を達成しやすいが、フォーマルなテキストで学習した情報抽出モデルをインフォーマルなテキストに適用すると精度が大幅に低下してしまう。このため、よりロバストな情報抽出モデルを開発するためには、インフォーマルなテキストを対象にした情報抽出のデータセット構築が必要となる。

本論文では、上記３つの要請に応えるべく、工場設備の保全レポートを対象に大規模かつ複数のサブタスクのラベルを付与したデータセットを構築する。我々のデータセットは約 1500 文書から構成されており、大規模であることと、4 つの情報抽出のサブタスクのラベルが付与されていることが特徴である。また、保全レポートはニュース記事と比較してインフォーマルな書かれ方が多く、情報抽出対象としてチャレンジングであることも特徴である。

2　関連研究

英語のマルチタスク情報抽出の研究でよく用いられるデータセットである ACE2005[1)] や SciERC[1] においては、サブタスクは固有表現認識と関係抽出の

1） https://catalog.ldc.upenn.edu/LDC2006T06

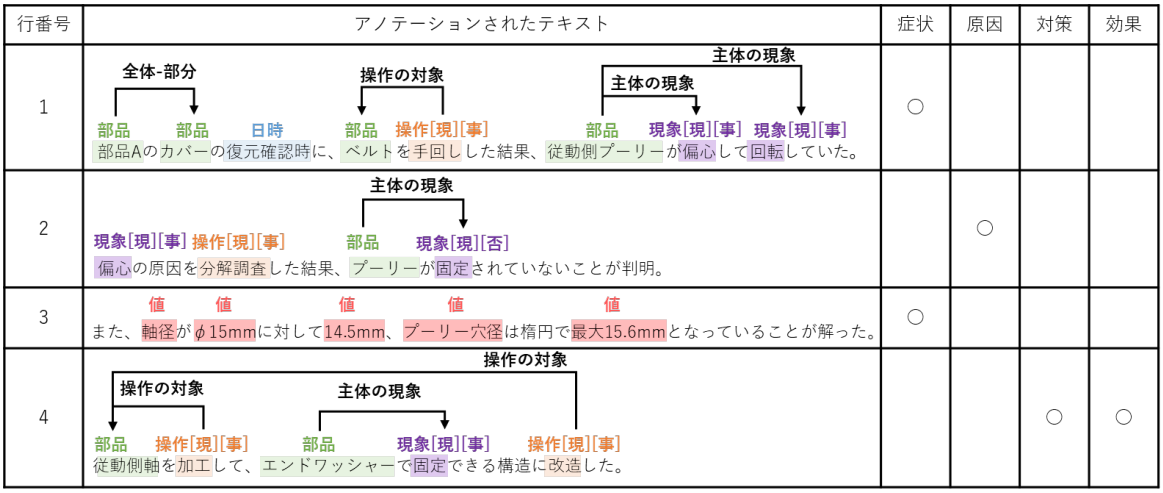

図 1　工場設備の保全レポートに対してアノテーションした例。テキスト中の固有表現の上部に固有表現ラベル、時制ラベル、事実性ラベルを示してあり、[現] は現在ラベル、[事] は事実ラベル、[否] は否定ラベルを表す。"症状"、"原因"、"対策"、"効果"の列は症状・原因・対策・効果文ラベルの有無を表す。

２つのみであり、また規模は約 500 文書と比較的小さい。日本語に限定すると、将棋の解説文を対象にデータセットを構築した研究 [2] があるが、規模は約 2000 文であり、固有表現認識とモダリティ推定の 2 つのサブタスクのアノテーションに留まっている。情報抽出の対象となるテキストとしてフォーマルではなく、よりノイズが多い Twitter を採用したデータセット [3] はあるものの、対象タスクは固有表現認識のみである。また、医療レポートを対象に固有表現認識、事実性判定、関係抽出の３つのサブタスクのラベルを 1156 文書にアノテーションしたデータセットを構築した研究 [4] がごく最近にあり、これが前述の３つの要請に最も近い既存のデータセットである。

3　データセット

本論文では、工場設備の保全レポートのテキストをアノテーション対象に採用する。富士フイルムエンジニアリングでは保全レポートに加えて設備情報やマニュアルなどの設備に関する情報を一元管理するシステムである KARTEMIX を開発している。保全レポートには工場設備に対する点検、メンテナンス、トラブル対応の活動がテキストで日々記録されており、大量のデータが蓄積されているため、大規模なデータセットが構築しやすい。設備やその状態などの表現が設備保全ドメインにおける固有表現とみなせるようにアノテーションの定義が容易である。また、固有表現間の関係や事実性判定など複数のサブタスクも設定できる。一方で、ニュース記事と比較すると、専門用語や略語が多く、箇条書きや名詞句の連続などの一般的な文よりは短い記述が見られるため抽出対象として難しい。また、下流のアプリケーションを考えた場合に、構造化情報の利活用のユースケースが想定しやすい。例えば、保全レポートからトラブルに関する情報を抽出し、構造化した形でデータベース化しておくことで、トラブル時に過去の対策や知見の検索や分析に活用することができる。本章では、アノテーション対象となるテキストとアノテーションの定義について説明する。

3.1　テキスト

保全レポートには様々な入力のフィールドが存在するが、"状況"、"処置"、"原因・対策"の３つのフィールドのテキストを対象とする。工場設備の実データから 1492 レポートを抽出し、アノテーションする。１レポートあたりの平均文数は 7.92 文であり、平均文字数は 259 文字であった。

3.2　アノテーション

本論文では保全レポートのテキストに対して、図 1 に示すような 4 つのタスクについてアノテーションを実施する。

3.2.1　症状・原因・対策・効果文ラベル

前述の通り、保全レポートには"状況"、"処置"、"原因・対策"の３つのフィールドが存在している。し

かし、実データを分析してみると、"状況"のフィールドにトラブルの原因や対策が記載されるように、必ずしもフィールド名の内容が記載されているわけではないことがわかった。このため、各文に対して、以下の４つのラベルをアノテーションする。

症状：設備の故障・不具合に関する状況を表した文（例：図 1 の文 1、文 3）。

原因：設備の故障・不具合が発生した要因・原因を表した文（例：図 1 の文 2）。

対策：設備の故障・不具合を解決するための措置を表した文（例：図 1 の文 4）。

効果：**対策**ラベルを付与した文のうち、設備の故障・不具合が実際に解決したことを表した文（例：図 1 の文 4）。ただし、対策を実施したが解決できなかったことが明示されている場合のみ**対策**ラベルを付与せずに、解決しなかったことが明示的に記載されていない文は**対策**ラベルを付与した。

１文に対して複数のラベルが該当する場合はマルチラベルをアノテーションする。

3.2.2 固有表現

工場設備のトラブルにおいて特徴的な固有表現を以下の 5 つに分類し、アノテーションする。

部品: 生産設備やその一部、もしくは生産に用いる材料を表す表現。例：送風機、モーター、ネジ

現象: 部品の動作が正常もしくは異常な状態を表す表現。例：異音、停止、漏れ

操作: 部品に対する操作や動作を表す表現。例：交換、調査、リセット

値: 測定値などの部品の状態を定量的に表す尺度やその値を表す表現。例：電流値、3mm

日時: 日時・時間を表す表現やタイミングを表す表現。例：先月、交換後、製造中

固有表現のスパンは名詞、数詞、記号の連続の最長範囲をアノテーションする。ただし、**操作**に関しては動詞の語幹も対象とした。ネストした固有表現は対象外とし、最長範囲のラベルのみを付与する。

3.2.3 モダリティ

前節で定義した**現象・操作**が付与された固有表現に対して、事実性ラベルと時制ラベルの二種類のモダリティを付与する。事実性は固有表現が表す事象が実際に発生したのかどうかを表すモダリティで、実際に発生した事象に対しては**事実**、発生していない事象に対しては**否定**をアノテーションする。確定的ではなく、推定を含むものも含めてアノテーションする。ただし、『これが原因だとすると交換が必要であるが』の"交換"のように仮定や条件など事象の発生の有無が明確ではない場合、**その他**を付与する。時制ラベルは発生した事象の時系列を表すモダリティでである。単純な時制表現で判断するのではなく、レポートのメイントピックとなっているトラブル・故障などのイベントに関する一連の事象をまとめて**現在**として、判断する。例えば、図 1 の文４における"改造"は過去を表す助動詞"た"が接続しているが、このレポートのトラブル対応の一連の活動に含まれる行動であるため、時制は**現在**となる。一方で、『前回の点検時にモーターを交換した』のように現在のイベントよりも過去に発生した**現象**や行った**操作**に対しては**過去**をアノテーションする。また、『ベアリングを交換予定』のように未来に行われる事象に対しては**未来**をアノテーションする。

3.2.4 関係

２つの固有表現（固有表現１を subject、固有表現２を object とする）の間に成り立つ関係に対して以下の 4 種類ラベルを付与する。

全体-部分: subject、object いずれも**部品**であり、object が subject を構成する一部である関係。

操作の対象: subject が**操作**、object が**部品**であり、object が subject の操作の対象となっている関係。

主体の現象: subject が**部品**、object が**現象**であり、subject が object の現象の主体となっている関係。

因果関係: subject が**操作**もしくは**現象**、object が**現象**であり、subject がある要因・原因を表し、その結果として object の事象を誘発することを表す関係。ただし、設備保全に関する専門知識がない作業者でも一貫性を持ってアノテーションを実施するために、"〜による"、"〜のため"のような手がかりとなる表現のパターンを 8 つ定義し、このパターンに合致し、因果が明らかな場合のみを付与対象とする。

4 データセットの統計量

本章ではアノテーションした各サブタスクのラベルの分布などの統計量を分析する。

4.1 文ラベルの分布

表 1 に症状・原因・対策・効果文ラベルの分布を示す。文ラベルとしては症状文が最も多く、順に対策文、原因文が多い。一般に不具合発生時にはその

症状を最も多く記載するため、これは直感に合致する。一方で原因よりも対策が多いことから、原因が究明できていなくても、対処療法的に行う一時的な対策を実施することも多いと考えられる。また、対策文のうち効果文である文の割合は約 67%であり、実際に効果があった対策が多く記載されていることがわかる。

表 1 症状・原因・対策・効果文ラベルの頻度分布

	文数	文書あたりの平均
症状文	2,786	1.87
原因文	1,507	1.01
対策文	1,987	1.33
効果文	1,333	0.89
ラベルなし文	5,860	3.93

4.2 固有表現ラベルの分布

表 2 に保全レポートの固有表現ラベルの分布を示す。出現頻度を比較して見ると、**部品**、**現象**、**操作**の順で多く、これは部品を主体として、その状態や部品に対する操作が保全レポートには多く書かれているという直感に合致する。文字長を見てみると、**部品**が最も長いことがわかる。これは**部品**の固有表現は『PC 電源ファン部』のように複数の部品で構成されることが多いためであると考えられる。

日本語で最も大規模な固有表現データセットである拡張固有表現タグ付きデータセット [5] と比較すると、1 文書あたりの固有表現数は 35.1 vs 29.4 であり、文字長は 259 vs 424 であり、我々のデータセットのほうが文書の長さが短いにも関わらず、固有表現の出現頻度が高いため、より多くの情報を構造化できていることがわかる。

表 2 固有表現ラベルごとの頻度分布

	用語数	平均文字長
時刻	4,792	4.20
値	5,712	4.61
部品	15,388	5.20
操作	11,906	3.80
現象	14,678	3.54
合計	52,476	4.17

4.3 モダリティラベルの分布

表 3 にモダリティラベルの分布を示す。事実性では**事実**が、時制では**現在**が最も多く、これは直感に合致する。一方で、それ以外のラベルが事実性では 17.2%、時制では 9.9%と一定数は存在しており、これは検索やデータ分析などの下流のアプリケーションにおいてはノイズになりうる。特に事実性の**事実**と**否定**は、『異音が発生』と『異音なし』のようにトラブルの事象として大きく異なるものを表しているため、区別が重要となる。

表 3 モダリティラベルの頻度分布

		表現数	文書あたりの平均
事実性	事実	21,643	7.77
	否定	2,524	0.91
	その他	1,822	0.65
時制	現在	2,3415	8.40
	過去	1,193	0.43
	未来	1,381	0.50

4.4 関係ラベルの分布

表 4 に関係ラベルの分布を示す。"関係あり"は関係ラベルが付与された固有表現ペア数、"関係なし"は固有表現のラベルの組み合わせとしては関係ラベルの付与対象であるが、テキストからその関係が認められなかった固有表現ペア数を表す。**因果関係**は他の関係と比較して、関係ありの頻度が非常に少なく、関係ありと関係なしの比率も最も小さい。これは今回のアノテーションでは手がかり表現から明示的に関係がわかるもののみを**因果関係**としてラベル付けしたためであると推定できる。手がかり表現のパターン数を増やす、もしくは手がかり表現がない暗黙的な関係も含めて付与対象とすることで、抽出対象を拡大することは今後の課題である。

表 4 関係ラベルの頻度分布

	関係あり	関係なし	関係あり/なし
全体-部分	3,211	12,872	0.25
操作の対象	5,473	4,845	1.13
主体の現象	8,686	11,663	0.74
因果関係	990	18,100	0.05
合計	18,360	47,480	0.39

5 おわりに

工場設備の保全レポートを対象に、大規模かつマルチタスクな情報抽出データセットを構築した。我々のデータセットの特徴は、インフォーマルな書かれ方が比較的多い保全レポートを対象に、文分類・固有表現認識・モダリティ推定・関係抽出の 4 つの情報抽出サブタスクのラベルが約 1500 文書に付与されていることである。また、データセットの統計量を分析するとことで、データセットの特徴や工場設備の保全レポートから情報抽出する価値について議論した。

商標

KARTEMIX は富士フイルムエンジニアリング株式会社の登録商標です。

参考文献

[1] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3219–3232, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[2] 亀甲博貴, 松吉俊, John Richardson, 牛久敦, 笹田鉄郎, 村脇有吾, 鶴岡慶雅, 森信介. 将棋解説文への固有表現・モダリティ情報アノテーション. 自然言語処理, Vol. 28, No. 3, pp. 847–873, 2021.

[3] Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 138–144, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[4] Fei Cheng, Shuntaro Yada, Ribeka Tanaka, Eiji Aramaki, and Sadao Kurohashi. Jamie: A pipeline japanese medical information extraction system, 2021.

[5] 橋本泰一, 中村俊一. 拡張固有表現タグ付きコーパスの構築. 言語処理学会 第 16 回年次大会 発表論文集, 2010.