事故事例構造化コーパスの構築

東明幸太 [1]　福岡康大 [2]　森辰則 [1]
伊藤拓海 [3]
[1] 横浜国立大学大学院 [2] 横浜国立大学 [3] 株式会社 IHI
tomei-kota-pc@ynu.jp　fukuoka-kodai-hr@ynu.jp　tmori@ynu.ac.jp
ito9762@ihi-g.com

概要

製造業の企業では現場からの問い合わせに対する回答を生成する際に過去の事故事例の活用が求められている。しかし、企業に蓄積されているものは構造化されていないプレインテキストであるため活用が難しい。本稿では、現場利用者の情報要求に応じて検索・要約をし、事故の流れを可視化して提示することで原因等をわかりやすく示すシステムを作成するために過去の事故事例を構造化することを目的とする。そのために、事故の流れを示すことができる注釈付け手法を提案し、実際にコーパスの作成を行った。また、事故事例に付与された注釈に基づいて、事故事例中の事象の系列を可視化するためのシステムの試作を行った。

1　はじめに

製造現場や建築現場などにおいて、現場で技術的な問題や事故・装置の不具合が発生した場合、現場作業員から担当部署に対してそれら技術的要素や事故・装置の不具合への対処方法について問い合わせが行われ、担当部署が過去の類似事例を参考にして回答を返し、その回答内容に従って問題解決を図る、というプロセスが取られることが多い。しかしそういった問い合わせへの返答に際して、現状では部署の担当者が問い合わせ文の読解・過去の事故不具合事例データベースからの前例の検索・回答文の作成といった作業を人手で行っており、非効率的だという現状がある。そのため、事故・不具合問い合わせ文書の解析・データベースからの類似事例の検索・回答文の生成といった作業を自動で行うことのできるシステムに対して需要が大きい。また、現場では発生した事故や不具合について記録を保存している場合も多く、そうして蓄積された過去の事故・不具合事例集を効果的に活用可能なシステムに対しても需要が大きい。本研究では事故・不具合問い合わせ文書の解析・類似事例の検索・回答文の生成といった作業を自動で行うシステムの実現を目標としている。特に、利用者の情報要求に応じて検索・要約をし、事故の流れを可視化して提示することで原因等をわかりやすく示すことを目指している。

しかし、企業に蓄積されているものは構造化されていないプレインテキストであるため、そのままでは上記システムの実現は難しい。上記の目標の実現には、原因を含む事故につながる事象の流れを表すことができるテキストの構造化と構造化テキストに基づく、事故内容の可視化が必要である。そのために、本稿では、どのように構造化を行えば良いかについて、事故事例のテキストを分析し、構造化のための注釈付けの枠組みを検討した。同枠組みを用いて、後述の『失敗知識データベース』の一部の事例に対し注釈付けを行い構造化されたテキストコーパスを整備した。さらに、その構造化されたテキストを入力として、可視化を行うシステムを試作することにより、提案する構造化手法の妥当性を検証した。テキスト構造化の自動化は今後の課題とする。

2　関連研究

これまで事象の流れを抽出する研究の一つとして、原因と結果の組を抽出する因果関係抽出が盛んに行われてきた。坂地ら [1] や佐藤ら [2] の研究では、決算短信、有価証券報告書などの金融情報テキストに関して因果関係抽出を行っている。手がかり表現を用いることで原因・結果を含む文を抽出し、原因・結果を含んでいると判定された文から原因と結果の対を抽出している。これらは金融分野のテキストに対して有効であることが確認されているが、我々が対象としている事故事例に対しての有効性は不明である。また、原因・結果のみの抽出であり、情報から得られる事象の流れを網羅していない。

既存の、事故・不具合事例文書の活用を目的とした先行研究として、大森らの研究 [3] がある。大森らは、事例文の中心となる部品や製品の情報提示を行うという観点でこの課題を捉え、事例文からの事故・不具合に関与する製品の記述の抽出、および因果関係抽出を通した事故・不具合の原因に関する記述の抽出というアプローチを通した事故・不具合事例集の活用を検討している。この手法では事例文から中核となる製品やその製品と因果関係を持つ出来事を表す「単語そのもの」を抽出することを目指しており、事故全体の流れに注目して文章を構造化しているわけではないため、事故全体の流れを把握することはできていない。

事故全体の構造化を目指している研究として、畑村らが作成した『失敗知識データベース』[4][5] がある。失敗知識データベースとは、失敗知識のデータ化・容易な伝達を目的として、失敗事例を分析して知識として活用できるデータベースを目指し開発されたものである。失敗知識データベースの事例は原因、対策などの項目に分けて記述されており、知識として活用できる形になっている。そのため構造化に大いに参考になる。その一方で、このデータベースと同じ知識を自動的に整備するためには、自動化を意識した分析や手法の提案が必要である。さらに、失敗知識データベースでは事故の詳しい流れが構造化されている項目は存在しない。

我々の研究では、事故・不具合問い合わせ文書の解析・類似事例の検索・回答文の生成といった作業を自動で行うことのできるシステムの実現を目標としている。その目標の達成に向けて、本論文では、事故事例の流れを構造化し、今後の事故事例の活用に貢献するための正解となるコーパスを作成する。

3 コーパスの設計

3.1 事故事例の構造分析と構造化の方針

まず一般的に、事故などの事象の流れは、ある「状態」において、ある「動作」が行われた結果、新しい「状態」が生じるということの繰り返しで表現できる。「状態」に注目すると「動作」は「状態変化」を起こすものとして位置付けられる。それらの前提から、本稿では事故などの事象の流れの記述は、ある時点での対象物群の有り様を記述した「状態」とその状態を変化させる動作を記述した「状態変化」の二種類の要素によって記述されているとする。

状態　ある時点での対象物群の有り様
例)「低圧タービンの外側から 3 段目の動翼 1 本が、車軸への取付部が折れて脱落していた」

状態変化　状態を変化させる事象
例)「タービン本体の設計時に想定されなかった異常振動が起きた」

これらは基本的に時系列順で記述されているが、そうではないものも多く、事故の流れを簡単に把握することは困難である。そのため、我々のコーパスでは、時系列に沿った順序関係を「状態」や「状態変化」の間の関係として付与することで、事故の流れを表現できるようにしている。また、事故事例文書は実際の事故の流れを表す記述である「客観的な記述」と事例文書の書き手が原因などを推測、判断して記述した「主観的な記述」で構成されている。そこで、我々のコーパスでは主観的な記述を「思考・判断」として注釈付けをし、これらを客観的な記述と区別する。そうすることで、書かれている記述が実際に起こったことであるのか、書き手の思考・判断の記述であるのかを分けて理解することができる。区別を明示するために、主観的な記述であるということがわかる「思考・判断」を基本要素に追加した。書き手が行う「思考・判断」の対象として「状態」(例えば、「〜となっていると考えらえる」) や、「状態変化」(例えば、「〜が起こったと推測される」) が現れる場合がある。

思考・判断　書き手の事故に対する主観的な思考や判断
例)「バランス調整不良に起因した振れ回りによる共振が原因であった」

「思考・判断」とみなす基準は、明示的に現れる判断表現に基づいて判断する。明示的に現れる判断表現としては、「原因であった」、「考えられる」、「原因は〜のためであった」などがある。

上の例では、「振れ回りによる共振が**原因であった**」の「**原因であった**」という記述が「原因の認定」という判断表現であるため、「思考・判断」であるとみなす。もし、「振れ回りによる共振が起こった」という記述であれば客観的な記述とする。

3.2 注釈付けのための規則の検討

以上で設計したコーパスを作成するために、テキストに注釈をつける。そのための注釈付けの規則を検討した。その結果を以下に示す。

- 「状態」「状態変化」「思考・判断」を基本要素として注釈付けするため、それぞれに対応するテキストを <s> (state), <t> (transition), <j> (judgement) タグで囲む。
- 時系列に沿った順序関係を「状態」や「状態変化」の基本要素間の関係として付与する。文書から順序関係が明確にわからないものは並列に扱う。時系列上で隣接していると判断される二つの基本要素において、時間的に前である基本要素を「始点」、後である基本要素を「終点」と呼ぶ。基本要素の時間方向を明示するため、時系列でつながる基本要素の「始点」の注釈に d 属性 (direction) を付与し、値”1” を与える。対応する「始点」と「終点」が組であることを明らかにするために、それぞれの id 属性に同じ値を与える。id 属性の値は、「始点」と「終点」の組が文書内で出現する順番により値”1”からはじまる通し番号とする。
- 同一の事象を表現している「状態」「状態変化」の記述部を共参照関係とし、その関係にある両注釈に、c_id 属性 (coreference id) を付与し、同じ属性値を与える (入れ子の場合は一番外側のみ)。事故事例文書では、異なる場所で同一の事象を表現するテキストが出現しうる。事象の流れを正確に把握するためには、それらが共参照関係にあることを記述する必要がある。
- 「状態変化」には人が意志を持って起こしたものとそうでないものがある。両者を区別するため、人が意志を持って起こした「状態変化」を「有意志」、そうでないものを「無意志」とし、「有意志」である「状態変化」には i 属性 (intention) を付与し、値”1” を与える。
- 「思考・判断」か否かは、明示的に現れる判断表現の有無に基づいて判断する。明示的に現れる判断表現としては、「原因であった」、「考えられる」、「原因は〜のためであった」などがある。
- 「思考・判断」に内包されている「状態」、「状態変化」が確実に起こったものではないと読み取れる表現のものがあるため、それらを区別する。「思考・判断」記述部が不確実と読み取れる記述がされているなら uncer 属性 (uncertainty) を付与し、値”1” を与える。
- タグで囲う範囲は述語までとする。

4 コーパス作成

「失敗知識データベース」の「事例概要（事故の内容・原因・対処がまとめて記述してある）」を用いてコーパス作成を行った。機械カテゴリの 210 件、失敗知識百選（失敗事例の中から国内外の典型的な事例を 100 例程度取り上げ、読みやすく記述）の 106 件を対象とした (両者で 4 件の重複があるため) 合計 312 件の事故の事例概要を対象とし、注釈付けを行い、コーパスを作成した。

4.1 注釈付け

注釈付けは第一著者、第二著者の二名で行った。ルールに従い人手で各事例の事例概要に注釈をつけた。その結果をもとに注釈が一致する場所は正解、一致しなかった箇所に関しては、議論を行い、注釈の付け直しを行った。図 1 の原文章に対する注釈付けの例として、結果を図 2 に示す。

原子力発電所で、タービンが損傷して自動停止し、続いて原子炉が自動停止した。タービンのカバーを外して内部点検したところ、低圧タービンの外側から3段目の動翼1本が、車軸への取付部が折れて脱落していた。同様の翼を検査したところ、840本中663本の動翼で損傷が発見された。試験運転中に、タービン本体の設計時に想定されなかった異常振動が起きたことによる金属疲労が原因と推定される。仮復旧での運転再開までに9ヶ月間を要した。

図 1 原文章（機械カテゴリ：5 番目のデータ）

<t id="1" d="1">原子力発電所で、タービンが損傷して自動停止し</t>、続いて<t id="1"><t id="2" d="1">原子炉が自動停止した</t></t>。<t id="2" i="1"><t id="3" d="1" i="1">タービンのカバーを外して内部点検した</t></t>ところ、<s id="3"><s id="4" d="1"><s id="7">低圧タービンの外側から3段目の動翼1本が、車軸への取付部が折れて脱落していた</s></s></s>。<t id="4" i="1"><t id="5" d="1" i="1">同様の翼を検査した</t></t>ところ、<t id="5"><t id="6" d="1">840本中663本の動翼で損傷が発見された</t></t>。<j id="7" d="1" uncer="1"><s id="8" d="1">試験運転</s>中に、<t id="8"><t id="9" d="1">タービン本体の設計時に想定されなかった異常振動が起きた</t></t>ことによる<s id="9">金属疲労</s>が原因と推定される</j>。<t id="6">仮復旧での運転再開までに9ヶ月間を要した</t>。

図 2 提案するタグセットで注釈付けを行った文章（機械カテゴリ：5 番目のデータ）

4.2 構造化されたテキストからダイヤグラムを生成するシステム

我々の研究の目的の一つが、事例文書の要約・可視化である。そこで 4.1 節で注釈付けされたテキストを適切に可視化できるかどうかを検証するために、テキストから図による可視化表現（以下、ダイヤグラム）を生成するシステムを試作した。注釈の種類とそれに付与される属性の組に応じて図 3 に示す通り色付けを行い、区別を容易にした。

事例の流れを把握するため、「思考・判断」に対しては、内包する「状態」、「状態変化」がある要素のみを図に表示した。「思考・判断」が内包する「状態」、「状態変化」のノードをさらに枠で囲み一つのノードとした。注釈付けの段階では、不確実とされたか否かは「思考・判断」の属性であったが、図では「思考・判断」の内包する「状態」、「状態変化」にも継承されるものとした。

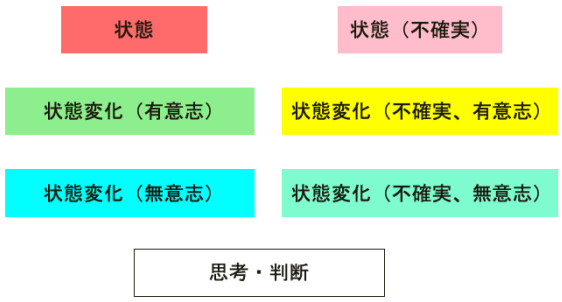

図 3　ノードの色と意味

実際に図 2 の注釈付けを行った事例に対して、テキストからダイヤグラムを生成するシステムを利用すると図 4 が出力される。ノードが有向エッジで結合され、時系列となるように配置されている。

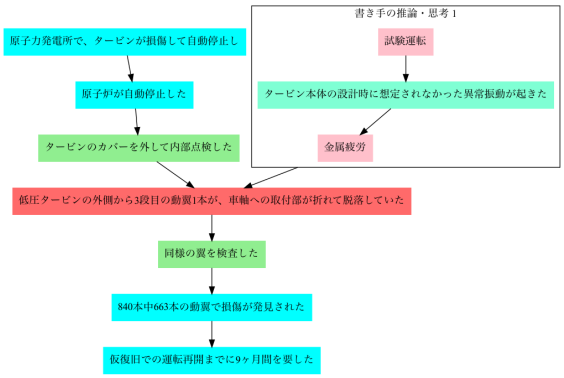

図 4　図 2 の事例からシステムが生成した図

5　考察

4 節で「失敗知識データベース」の「事例概要」316 件に対して、流れを把握できるような構造化のルールを作成し注釈付けを行った。注釈付け前後のテキストや生成されたダイヤグラムを第一著者、第二著者の二名が観察し、議論をした。図 4 のような可視化の結果により「動翼 1 本が、車軸への取付部が折れて脱落していた」という状態が「鍵となる状態」であることがわかりやすくなった。それに対する「試験運転による想定外の異常振動が起き、金属疲労状態になったであろう」という書き手の推論から「動翼 1 本が、車軸への取付部が折れて脱落していた」という原因状態に至る過程の推測もわかった。注釈付けを行っていく中で注釈の位置やどの注釈にするかを修正をしながらコーパスを作成した。特にタグセット作成後、どのタグを振るかという検討する際に判断基準を明確にするべきものがあったため以下に示す。

「状態」「状態変化」の違い　「エア吸気口に水滴となって残り」の部分に「状態」、「状態変化」どちらのタグを振るかについての検討の余地があった。検討の結果「残り」は動作そのものなので「状態変化」であり、「残っている」であれば進行を表す相表現が付加されているので「状態」と判断する。具体的には、「〜ている」、「〜ていた」といった進行を表す相表現があれば、その文を「状態」と判断する。

動作性名詞の扱い　動作性名詞とは、例えば、「レールの熱膨張で...」における、「熱膨張」のように、サ変名詞や動詞由来の名詞である。動作を名詞化したものであるから、「状態変化」とする。

6　まとめ

本稿では、プレインテキストの活用のための事故事例の構造化のために事故事例テキストを分析し、構造化のための注釈付けの枠組みを検討した。構造化ルールを作成し、注釈付けコーパス作成を行った。さらに、構造化されたテキストを入力として、可視化を行うシステムを試作することにより、提案する構造化手法の妥当性を検証した。正解データとして 312 件は多くないと考えられるため、他の事例に対しても注釈付けを続けていく必要があると考えられる。その後、今回人手で行った構造化を自動化することで、事故事例を把握する時間が短縮されることに貢献できることや、事故の真の原因特定や回答文生成にも役立つと考えられる。

参考文献

[1] 坂地泰紀, 酒井浩之, 増山繁. 決算短信 pdf からの原因・結果表現の抽出. 電子情報通信学会論文誌 D, Vol. 98, No. 5, pp. 811–822, 2015.

[2] 佐藤史仁, 佐久間洋明, 小寺俊哉, 田中良典, 坂地泰紀, 和泉潔. 有価証券報告書からの因果関係文の抽出. 人工知能学会全国大会論文集 第 32 回 (2018), pp. 2O404–2O404. 一般社団法人 人工知能学会, 2018.

[3] 大森信行, 森辰則. 不具合事例文からの製品・部品を示す語の抽出ー語の実体性による分類ー. 電子情報通信学会論文誌 D, Vol. 95, No. 3, pp. 697–706, 2012.

[4] 畑村洋太郎, 中尾政之, 飯野謙次ほか. 失敗知識データベース構築の試み. 情報処理, Vol. 44, No. 7, pp. 733–739, 2003.

[5] 失敗知識データベース, （2022-10 閲覧）. http://www.shippai.org/fkd/index.php.