情 報 学 基 礎 61−20
自 然 言 語 処 理 142−20
（2001. 3. 6）

障害情報からのマイニング

斉藤孝広† 渡部勇†
{takahiro,isamu}@flab.fujitsu.co.jp
† 富士通研究所
† 〒211 川崎市中原区上小田中 1015

テキストマイニング技術を用いて、障害管理情報からの有用な知識を発見する方式について提案する。障害管理情報とは、製品試験の際に発生した障害、その原因・対策を自然文で記述したものであり、通常は発生時の環境などコード化可能な情報と合わせて1エントリとして管理される情報である。

本方式は、この障害管理情報の障害・原因・対策の各自然文記述部分より、その記述内容を端的に表現する情報単位を係り受け解析を用いて抽出し、それらの間の関連性をマップとしてユーザに提示する事により、ユーザの分析作業を支援するものである。

The mining from trouble control data

SAITO Takahiro† WATANABE Isamu†

† Fujitsu Laboratories Ltd.
† 1015, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211 Japan

This paper describes a formula and a tool to support the user to get valuable knowledges in the trouble control data with the “text mining” technology.

One trouble control datum describes one troulbe happend in product test with texts which describe a state of the trouble,a cause and a countermeasure for it.

The tool extracts the “outline” from each text using the informarions of the dependency structure in it first,then exhibits the dependencys among the “outline”s with a map to the user.

In this paper, we illustrate the instance of valuable knowledges acquired from trouble control data with this tools.

1 はじめに

電子化されたテキスト情報を有効に活用するための技術として、その情報源の傾向など、個々のデータを見ただけでは得られない種類の情報を抽出するテキストマイニングが注目されている [1]。このテキストマイニング技術は、例えばコールセンタに寄せられる顧客の声や営業日報などの大量のテキスト情報に対する分析に適用されている [2],[3]。

一方、筆者らは、テキストマイニング技術の適用対象として、製品開発に伴う製品試験の際に作成される障害管理情報に注目した。障害管理情報とは、試作した製品を実際の使用状況に合わせた種々の環境で製品を動作させ、その際に見つかった不具合を、その原因、対策、発現環境などの情報と合わせて1つの障害事例としたものである。これらをまとめた障害管理表には、個々の事例を検索すれば獲得可能な知識だけでなく、種々の障害に対する原因および対策の全体的な関連を見る事で初めて分かる有用な知識 (例えば、「ある障害について、他の障害を含めて考えた時に最も有効な対策は～である」) も含まれており、これらの知識を獲得するためには分析作業が不可欠となる。これらの知識は設計・開発部門にフィードバックされ、製品開発に有効に活用されるべきものであるが、多量の障害事例における障害・原因・対策の各項目の関係は複雑であり、その分析作業自身のコストが高いために充分に行なわれていないのが実情である。本稿はこの分析作業をツールにより支援する方式について報告を行なうものである。

2 分析手法

2.1 情報抽出を用いた分析手法

筆者らは、ACCENT[4] とよばれるテキストマイニングツールを開発している。ACCENT は、分析対象とするテキスト文書群の中で用いられている単語の間の関連度を共起性に基づいて計算し、それらを概念マップとして可視化することで、ユーザの分析作業を支援するものである。

このツールを本稿における分析目的すなわち障害内容と原因および対策の間の関連性を把握するために適用する事を考える。まず、各々の項目の記述内容がキーワードのみである場合や、自然文による記述であってもその内容を1語で表現するキーワードが文中に含まれている場合には、現 ACCENT の枠組内でそれらのキーワード間の関連を見ることができるのでそのまま適用可能である。

しかしながら、一般の障害管理情報においては、この仮定を設ける事はできない。その内容を一つのキーワードで表現できる場合でも、その表現のバリエーションは無数にあるからである。このような障害管理情報を ACCENT で扱うためには、記述文よりその項目の内容をよく表し、かつ表現のバリエーションを吸収した情報 (以後、この情報を「概要情報」と呼ぶ) を処理単位 (つまり単語) とする事が必要となる。

そこで本手法においては、この概要情報を情報抽出技術を用いて抽出する方式を採用した。このような情報の抽出方式としては他にも、項目毎に予め分類カテゴリとそのラベルを設定しておき、記述文を自動分類技術を用いて分類して概要情報としてそのカテゴリのラベルを用いるという方式も考えられるが、この方式には以下の問題がある。

- 予め分類カテゴリを用意する必要がある。

  また、予期していなかったカテゴリには対応できない。といった問題も生じる。

- 適切な抽象化レベルを予め設定できない。

  各障害事例を分類し、そのカテゴリのラベルで内容を代表させるということは記述内容の抽象化を行なうのに他ならない。その際の抽象化のレベルは分類すべきデータに依存するのでそれを予め設定するのは不可能である。

またこれらの抽出結果間の関連は、「障害とその原因」といった明確な意味付けが与えられるので、ACCENT が作成するマップもその意味付けを直接的に反映した表現方式を採用した。

まず、本手法に用いられた情報抽出方式及び可視化方式について簡単な説明を行なう。また、次節において概要情報抽出手法の詳細を抽出実験の結果と共に報告する。

2.2 概要情報抽出方式

筆者らは、一般的な障害管理情報における記述文においては、その内容を推察することが可能な

最小単位が係り受け組であると考えた。そこで、記述文中の係り受け組について、その項目を記述している係り受け組を一つまたは少数選択し、それらから決まったルールに基づいて文字列を生成し、それを概要情報とした。

本方式では、以下の3つのルールを用いて抽出を行なう事にする。

1. 抽出文選択ルール

   記述が複数の文からなる場合に、文の重要度を評価して抽出を行なう文を選択するルールである。

   例えば、プリンタにおける障害管理情報の障害記述文「連続印刷において黒ベタ部がかすれる。1000枚中30枚程度に発生。」については、ルールにより最初の文、「連続印刷において黒ベタ部がかすれる。」を選択する。

2. 係り受け組選択ルール

   選択された文について係り受け解析を行ない、得られた係り受け組の重要度を評価して概要情報に盛り込むべきものを選択するルールである。

   上の例においては、係り受け解析の結果、まず以下の係り受け組を得る

   連続印刷において - かすれる
   黒ベタ部が - かすれる

   ここで得られた係り受け組に対して、重要度の評価を行ない、「黒ベタ部が-かすれる」という係り受け組のみを選択する。

3. 概要情報構成ルール

   選択された係り受け組のセットより概要情報とする文字列を作成するためのルールである。

   上の例なら、選択された係り受け組より「黒ベタ部カスレ」を作成し、これを概要情報として抽出する。

なお、本方式は係り受け解析を用いている。今回使用した係り受け解析エンジンは、以下の処理を行なって係り受け組を抽出するものである。

1. 入力文を形態素解析する。
2. 形態素解析結果を、ルールに基づいて文節に分割する。
3. 文節内の構成形態素によって、各文節の素性を付与する。
4. 各文節間の係り受け確率を素性より算出し、以下の制約下において全ての係り受け確率の積が最大となるような係り受けパターンを出力する。
   - 最後の文節を除き、各文節はそれより後方に必ず一つの係り先文節を持つ。
   - 係り受け関係の交差はない。

このエンジンにおいてその解析精度を決定するのは、素性の選択とそれに依存する係り受け確率の算出方式である。今回は、使用した形態素解析エンジンの出力結果より判定可能なもののみを素性とし、それに基づく係り受け確率を人手で設定した。この係り受け確率を自動学習によって設定する[5]事で、解析精度が向上する可能性はあるが、1) パラメータとなる素性の数自身が少ないのでその解析精度はすぐに頭打ちになる可能性が高い、2) 次節の抽出実験においては係り受け解析の精度が抽出精度に大きな影響を与えていない[1]、という理由により自動学習は用いていない。

2.3 可視化方式

概要情報間の関連は、「障害内容とその原因」、「ある原因に対する対策」など、非常に明確なものとして定義可能である。従って、これらの関連の意味を反映した可視化方式が有効である。

そこで今回は、ACCENTに組み込まれているD-ABDUCTOR[6]の階層レイアウトと呼ばれる機能を利用した階層マップを可視化情報の提示に用いる事にした。

階層マップとは、2つの単語グループに対して、各々のグループを一つの階層とし、1つのグループ内の各単語に関してそれに関連する別グループの単語とを線で結んだものである。階層マップ内の各単語は、階層レイアウトによって線の交差が

[1] これについては次節で考察する

最も少なくなるように配置されており、その間の関係が見やすくなっている。また、その関連度の強さを線の太さで表現する事で、関連度という数値的情報も図の形で表示することができる。

特に、3つの単語グループA,B,Cに対して、A-B,B-Cといった組の各々を階層レイアウトで配置した階層マップを2段階マップと呼ぶ事にする。

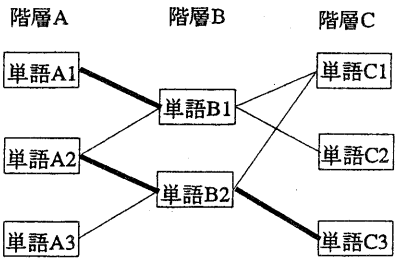

図 1: 階層マップ (2段階マップ) 例

この階層マップにおいては、「障害」、「原因」、「対策」と言った各項目における概要情報が、各々別階層に表示され、かつ各々の階層間を結ぶ線の意味が階層毎に決まっているので、各項目間の関連度を整理して分析者に提示するといった目的に非常に都合が良い。また、階層マップにおいてはその関連度を線上に表示することができるので、関連度の強さを単純に共起頻度とすることで、線上にその件数を表示することができる。例えば、「障害」グループ内のある単語と、「原因」グループ内のある単語の間の関連度が1であった場合、その原因でその障害が起こった事例が1件ある事を意味することになる。

3 概要情報抽出

まず、本手法による概要情報の抽出ルールとその評価実験結果について述べる。次に、抽出した概要情報を分析に用いる際に考慮すべき点を挙げてそれに対する今回の対応方法を述べる。

3.1 抽出ルールと評価実験結果

本手法における概要情報の抽出は、文選択ルール、係り受け組選択ルール、概要情報作成ルールの3種類のルールを入力テキストに次々に適用して行なう。

以下にそれらのルールの詳細な説明を行なう。例として挙げられているルールは、まずあるプリンタに関する障害管理表における抽出ルールを人手で作成し、次に作成したルールについて汎用性(他の障害管理表でも正しく機能するか)を検討し、汎用的と思われるもののみを採用したものである。

これらのルールは項目別に設定されるが、一つの障害事例においては、「障害」の概要情報が1つしか抽出されないのに対し、「原因」「対策」に関しては、複数個が抽出される可能性がある。これは、複数の障害が一つの事例に記述されることはないが、ある一つの障害の原因が複数あり、その各々に対して対策を行なうといった事は普通に行なわれている事を反映している。この複数個の抽出結果の扱いについては、後に述べる事にする。

1. 文選択ルール

   - 「障害」: 最初の文
   - 「原因」: 全ての文
   - 「対策」: 全ての文

   ただし、ここで「文」は2つ以上の文節を持つものとしている。1つしか文節を持たない文、つまり、キーワードのみが書かれているものは、それが発現環境などの情報を記述したものであり、項目を表現するキーワードではないとしている。

2. 係り受け組選択ルール

   係り受け組選択ルールは、選択した文で表現されている内容の内、概要情報に盛り込むものを選択するルールである。例えば、「障害」における概要情報としては「〜が」「〜した」という要素は最低限必要であり、一方、「原因」においては「〜が」「〜である(ため)」、「対策」においては「〜を」「〜した」という要素が必須であるので、このような情報が含まれている係り受け組を選択するルールである。

   日本語の構造上、このような情報を抽出するためには、文の述語に注目すれば良いはずで

ある。従って、ルールとしては、文中の述語文節を判定にその文節を係り先とする係り受け組の中で、概要情報に盛り込むべきものを選択することになる。また、「原因」「対策」においては、例えば原因なら「AがBし、CがDしたため。」といった、複数個の原因を一文で記述している場合も少なくないが、このような並列構造を認識するのは難しいので、動詞を含む文節全てについて、抽出を行なう事にした。

また、動詞を含む文節だけをキーとするだけでは不十分で、例えば、「ドラム内部の異物のため、～」といった、動詞文節が省略されている表現(意味的には「ドラム内部に異物が混入したため、～」)もある。このような表現からの抽出を行なうために、「あるキーワードに係る係り受け組」も抽出対象としている。このキーワードは例えば上の例における「ため」である。

このような係り受け組選択ルールで、一つの概要情報を作成する際に用いられる係り受け組の個数を、1～2個に制限する。

3. 概要情報作成ルール

係り受け組選択ルールによって、概要情報として盛り込む内容が選択されるが、それだけでは、表現のバリエーションの吸収処理は行なわれていない。この表現の統一化処理を行ない、最終的な概要情報とする文字列を作成するのが概要情報作成ルールである。

例えば、「印字がかすれる」という障害の記述としては、「印刷が掠れる」といった単純な同義語や表記揺れの違い以外にも、「印字カスレが発生」、「印字にカスレが生じる」という文の構造自身も異なる表現が可能である。これらを統一化するため、まず、選択した係り受け組の文節に対して以下の処理を行なう。

- 助詞の削除
  例えば、「裏面に横スジが入る」、「裏面で横スジ発生」という二つの記述文における「裏面で/に」という揺れを吸収するために、助詞を削除する。
- 品詞の変換
  述語となる動詞を名詞形に変換する。例えば、「かすれる」を「かすれ」に変換する事で、文の構造の異なる表現の統一化を行なう。また、その動詞文節が否定形である場合、後ろに「不能」を付与する事で表現する。
- 重要度の低い動詞の削除
  例えば、「障害」においては元々が障害が発生した際にその内容を記述するものであるので、「発生する」という述語の「障害」の概要情報における重要度は低い。このような動詞を、概要情報から削除する。これにより概要情報「～発生」が「～」のみとなる。このような動詞は、他にも「障害」における「ある」「入る」、「原因」における「よる」、「なる」、対策における「行なう」などがある。
- 同義語／表記の統一
  例えば「印刷」と「印字」といった、同義語の統一を行なう。ただ、これらの同義語が全ての障害管理表の記述において成立するかどうかは不明であり、今回は汎用的ルールのみを選択する事にしたので、このような同義語変換ルールは一切採用していない。また、表記の統一に関しては、抽出結果中で、平仮名のみで構成される単語はすべてカタカナに統一することにした。

次に各文節の処理結果を接続する。その際に語順の統一を図るために、係り受け組選択ルールで必須とされた係り受け組からの処理結果を結合し、それに意味を追加する様な係り受け組も選択された場合にはその要素をその前に付与することで概要情報となる文字列を作成する。例えば、「黒い縦スジが発生した。」より「黒い-縦スジが」「縦スジが-発生した」という係り受け組が得られ、ルールによって必須の係り受け組が後者、その意味を強化する係り受け組として前者が選択されたばあい、各文節の処理結果は各々、「黒」、「縦

スジ」、ϕ (処理の結果、全て除去された) であり、概要情報として「黒縦スジ」を抽出する。

次にこのルールを別のプリンタに関する障害管理表に適用して抽出実験を行ない、結果をまとめたのが表1である。正解の判定は、その抽出結果だけで記述内容が推察でき、かつその内容が項目と合致しているものを正解とした。

	「障害」	「原因」	「対策」
正解率 (%)	87.8	70.1	70.3

表 1: 概要情報抽出結果

結果として、「原因」「対策」の正解率が低くなった。これは、「原因」「対策」に関しては、複数個の抽出を許しているため、その分ルール自身の制約が弱く、概要情報として全く意味をなさない単なるゴミが多く混じっているのが主な原因であるが、それ以外にも次のような抽出エラーがあった。

- 「原因」
  - 条件を表す記述からの抽出結果 (例えば「装置が停止した際に」より抽出される「装置停止」)
  - 「障害」の内容の繰り返し (例えば「～しており、裏面が汚れる」より抽出される、「障害」の概要情報と同一の「裏面汚れ」)
- 「対策」
  - 特定の言い回しによるノイズ (例えば、「～する様に変更する」より抽出される「様変更」)
  - 「原因」項目の内容を受けているため単独では意味不明なもの (例えば、原因が「バネ強度不足」である時に、対策として「～したものに変更する」より抽出される「モノ変更」)、
  - 対策が「現状通り[2]」である場合に、その根拠を述べた記述からの抽出結果 (例えば、「通常使用においては問題は発生しないので、～」より抽出される「問題発生不能」)

[2]通常使用においては殆んど問題にならないので特に対策を施さない事を意味する

これらを除去するには、各項目における抽出ルールの充実させるだけでなく、他の項目の抽出結果を考慮した処理が必要となるはずである。

また、本手法は係り受け解析を用いているので、その精度の向上も抽出精度に影響するはずである。しかしながら、今回の評価データにおいては、1) 入力文が簡潔な文である、2) 解析を誤りやすい係り受け組 (例えば、副詞の係り先) は係り受け組選択ルールでは選ばれないので概要情報の抽出精度には影響しない、といった理由より係り受け解析の精度は殆んど問題にはならなかった。

3.2 概要情報の抽象度の問題

概要情報の抽出目的は、表現のバリエーションを吸収して同じ内容を同じものとして集計するためのものである。

しかしながら、本当は同じ障害であっても、事例入力者が記述する内容自身に洩れがあったり、その表現の違いにより抽出結果が異なってしまう場合がある。後者の例としては、ルールの不備が原因であるものもあるが、例えば「障害」からの抽出結果が各々「53mm 周期横スジ」、「ドラム周期横スジ」である二つの障害事例において、この製品におけるドラム周期が 53mm であるといった特有の知識があってはじめてこの障害事例は同じ障害であると判定可能なものもある。

この問題を解決するためには抽象化を行なう必要がある。上の例であれば、2つの抽出結果の共通部分文字列「周期横スジ」を集計単位とすることである。どの程度の抽象化を行なうかは、他の抽出結果との兼ね合いで決定する事ができ、例えば [7] の方式が適用可能である。

しかしながら、このような抽象化を行なう事で、障害に関する原因究明の際に有用な弁別情報が失われてしまう可能性がある。例えば、様々な周期の横スジが発生する障害の中で、その周期がドラム周期である場合は、その原因が「ドラム傷」と一つに特定できる場合、「ドラム周期」は「周期横スジ」障害における原因究明の際に有用な弁別情報である。これらの有用な情報を除去して全て

の周期の横スジ障害を「周期横スジ」として抽象化を行うのは、分析ツールとしては問題があると思われる。

以上より、「障害」に対しては、抽出結果を意味的まとまりに分割して、その全ての後方部分文字列を抽出結果として集計を行なう事にする。つまり、抽出した概要情報が「53mm周期横スジ」であるならば、「周期横スジ」、「横スジ」、「スジ」といった後方部分列に対しても概要情報として集計を行なう事にした。これにより、一つの事例が種々の抽象度を持つものとして分析を行なう事ができる。

なお、「原因」「対策」については今回の場合は、このような問題は見られなかったので、今回はそのまま集計を行なった。

3.3 複数抽出の問題

今回作成した抽出ルールにおいては、「原因」、「対策」については、複数個の概要抽出を許している。これは、「原因」ならば複数個の原因があってはじめて発現するような障害を表現する事を想定しているからである。

しかしながら、実際に抽出結果を評価してみると、例えば「プラップが変形し、ローラーと接触する」から抽出される二つの概要情報「プラップ変形」「ローラー接触」といったような、他方がもう一方の根本原因となっている場合が殆んどであった。この場合は、どちらかを選択すべきであるかもしれない。

ただ、抽出した複数個の原因が、各々独立であるか、このような依存関係にあるかを判定するのは、今回の方式では非常に難しく、また、事例によってはそのどちらか一方しか記述されていないものや、挙げられた原因の各々に対して対策をとっている事例もある。

そこで今回は、複数個の抽出結果を各々別々のものとして集計することにする。さらに「対策」からも複数の概要情報が得られた場合には、各々別に関連を付与する。例えば、一つの障害事例より「障害」概要情報が一つ(これを F とする)、「原因」概要情報が2つ(各々 C_1, C_2 とする)、「対策」概要情報が2つ(各々 M_1, M_2 とする)抽出された場合、「障害」-「原因」関係としては $F-C_1, F-C_2$ の2つが、「原因」-「対策」関係としては、$C_1-M_1, C_1-M_2, C_2-M_1, C_2-M_2$ の4つが集計されることになる。

これにより、例えばある原因に対する対策が、別の原因と結ばれてしまうといった問題が起こってしまうが、その対応関係を自動的に判定するのは難しい。そこで、今回はその判断を分析者に委ね、ACCENTのマップ編集機能を用いて問題部分を人手で解決する事にした。なお、この作業の際に必要な情報は、マップから元データを検索する機能で効率的に収集可能である。

4 分析例

ACCENTを用いて、本手法による実際の分析作業の事例を挙げる。

図2は、「障害」-「原因」-「対策」の間の関連度を2段階マップで表したものの一部である。この図は以下の操作によって作成された。

1. (マップの作成)

   「障害」-「原因」-「対策」の2段階マップを作成

2. (「原因」の編集)

   一つの「障害」と複数の「原因」が結ばれている場合、その因果関係を判定する。因果関係があった場合、その「対策」も考慮し適当なものを残して削除する。

3. (「対策」の編集)

   一つの「原因」と複数の「対策」が結ばれている場合、本当にその原因の対策となっているもののみ以外の線を削除する。

このマップより、以下が読みとれる。これは、個々の障害事例を独立に見ている限りにおいては得られない知識であり、本方式の有用性が確認できる。

- 障害「白ベタ部汚れ」と「定着不良」は、共通の原因を持つ。
- その原因の一つである「ニップ減少」に対する対策は「用紙の間を広げる」事である。この対策は「用紙下端シワ」にも有効である。

- もう一つの原因である「定着温度不足」に関しては「170℃一定制御」が有効である。この対策により定着温度が上がるが、あまり上げ過ぎると「用紙下端シワ」という別障害を招くことになる。

- 以上より、マップ内の3つの障害を解決するには、「用紙間を広げて定着温度を適切に設定する」事が必要である。

5 まとめと今後の課題

以上により、障害管理情報を整理した形でユーザに提示し、ユーザの分析作業を支援する方式について、実例を通してその有効性を示した。

今後の課題としては、個々の項目からの概要情報の抽出精度の向上は勿論であるが、一方、より汎用的なルールの作成も重要であると考えている。今回は人間の判断でルールの選択したが、種々の障害管理表において自動学習により汎用的なルールを作成する機構なども検討すべきである。また、前節で行なったような、人手による編集作業を自動化できればより有用なツールとなりえるが、それには記述文の内容により深く踏み込んだ自然言語処理と、他の項目からの抽出結果を考慮した大局的処理の二つが必要となる。

参考文献

[1] 那須川哲哉 他: テキストマイニング - 膨大な文書データの自動分析による知識発見 -, 情報処理,Vol40 No.4,pp358-364,1999

[2] 長野徹: テキストマイニングのための情報抽出, 情報処理学会研究報告,FI-60,pp31-38,2000

[3] 市村由実 他: 営業日報を対象としたテキストマイニング - 成功事例および機会損失情報の抽出 -, 第14回人工知能学会全国大会, pp532-534,2000

[4] 渡部勇 他: 単語の連想関係によるテキストマイニング, 情報処理学会研究報告,FI-55,pp57-64,1999

[5] 内元清貴 他: 最大エントロピー法に基づくモデルを用いた日本語係り受け解析, 情報処理学会論文誌,Vol.40 No.9,pp3397-3407

[6] 三末和男 他: 図的発想支援システム D-ABDUCTOR の開発について, 情報処理学会論文誌,Vol.35 No.9,pp1739-1749,1994

[7] 斉藤孝広 他: 連想検索における属性語の抽出方式, 第14回人工知能学会全国大会,pp171-172,2000

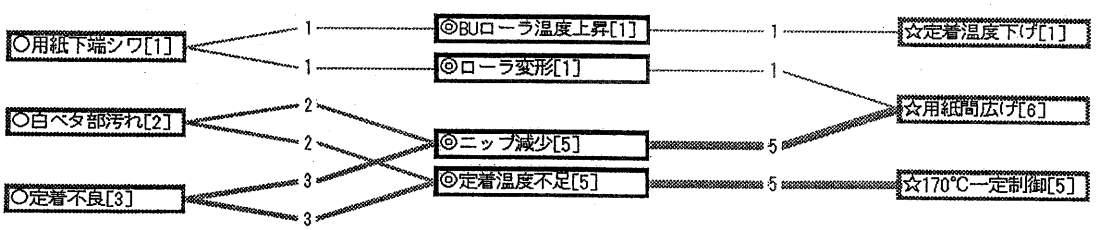

図 2: 「障害」-「原因」-「対策」の関連マップ