航空安全情報からのトラブル発生パターンの抽出について

齊藤孝広† 渡部勇† 松井くにお† 寺田昭‡ 齋藤隆‡
{saitou.takah-02,watanabe.isamu,matsui.kunio}@jp.fujitsu.com
{akira.terada,takashi.saito}@jal.com
† (株) 富士通研究所, ‡ (株) 日本航空インターナショナル

1 はじめに

航空会社においては、安全運航は至上命題である。同種の事例の防止には、ヒヤリハット・インシデント等の情報の活用が重要である。この情報を本稿では航空安全レポートと呼ぶこととするが、その内容は定型項目と自然文による詳細記述で構成されている。航空安全レポートは1つの事例に対する対応を検討する際に利用されるだけでなく、これらをまとめて分析を行なうことで総合的な対応を検討する際にも利用される。

一方、多量の文書データから主な記述内容を把握することはテキスト分析のタスクとして一般的である。このタスクを支援する手法としては文書クラスタリングが挙げられる。クラスタリング処理により作成されたクラスタを解釈する手段としては、クラスタを特徴付けるキーワード (特徴キーワード) を提示する方法が一般的であるが、それだけではしばしば解釈の難しいクラスタが作成され、特にクラスタ中に複数の内容が含まれる場合に大きな問題となる。

そこで本稿では、文書クラスタの内容を解釈する作業を支援する方式として、キーワード間の共起関係を利用する手法を提案する。また、本手法を航空安全レポートを分析対象として、その中の「乗員・乗客の負傷を伴う事例」の主な発生パターンを抽出するタスクに適用した実験についても報告する。

2 分析方式

本稿の提案する分析方式の基本的な流れは以下の通りである。

1. 前処理
2. 部分集合への分割

   分析対象とする文書群をクラスタリングにより部分集合に分割する。
3. 特徴キーワードの抽出

   作成した部分集合を特徴付ける特徴キーワードを抽出する。
4. 内容推定

   特徴キーワードから文書の主な記述内容を推定する。その際、特徴キーワード間の共起関係に着目して記述内容の分離を行なうことで、分析者の推定作業を支援する。

次節で各フェーズの処理内容について述べる。

2.1 前処理

今回の分析対象である航空安全レポートは、専門用語の略語が非常に多く含まれており、同じ意味であってもその表記に多くの揺れがある。このままでは、検索処理やクラスタリング処理の精度に悪影響を与える可能性が高いので、用語の統合化ルールを作成した。このルールは、いわゆる同義語の言い換えだけでなく、その上位概念語への言い換え (例えば「CAT(Clear Air Turbulence の略語)」 ⇒ 「Turbulence」) も記載することで、検索の際の再現率やクラスタリング精度を向上させることができる。

今回は、統合ルールを専門用語を重点に整備し、1768 個のルールを作成した。作成したルールにより、例えば今回の分析対象データにおいて「Turbulence」で検索したときの結果は 54 文書から 186 文書まで増加するが、多義語を原因とする過剰な統合が行なわれた文書が 15 文書あった[1]。今回の分析目的においては、この程度の過剰統合は許容

[1] これらは全て「CAT」が「カテゴリ」の意味で使用されていた。

できると考え、高度な処理が必要となる多義性の解決は行なわないことにした。

2.2 部分集合への分割

文書中のキーワードを抽出[2]して特徴素とし、一般的な階層クラスタリングアルゴリズムを用いてクラスタリングを行なった。

2.3 特徴キーワードの抽出

まず、集合中の文書に出現するキーワードを χ^2 検定 (棄却限界 1%) し、集合と正の相関のあるキーワードを特徴キーワードとして抽出する。ただし、ここで計算される χ^2 値はキーワードと集合との相関の強さを表す値であり、出現頻度に関する情報が落ちる。一方、今回の目的は集合内の文書の主な内容を抽出することにあるので、クラスタの特徴をよく表し、かつ出現頻度の大きい特徴キーワードを見つける必要がある。そこで、キーワードの集合中の出現頻度と χ^2 値の積をキーワード評価値として採用し、この値でランキングする方式を採用した。

2.4 内容推定

特徴キーワードから集合内の文書に記載されている主な内容を推定する作業は、集合内に複数の内容が含まれ、複数の内容から特徴キーワードが抽出されている場合に、より困難となる。そこで、抽出されたキーワードを内容毎にグルーピングすることにより、この推定作業を支援する。

キーワード間のグループ化方式として、キーワード w_1, w_2 に関して次の式で定義する共起度を算出し、その値の大きなキーワード組でグラフを作成する方式を採用する[3]。

$$\begin{aligned} s(w_1, w_2) &= P(w_2|w_1)/P(w_2) \\ &\quad + P(w_1|w_2)/P(w_1) \end{aligned}$$

ここで $P(w)$ は全文書における w の出現確率 (= w が出現する文書数 / 全文書数)、$P(\phi|\varphi)$ は φ が出現する文書に ϕ が出現する確率 (= φ 及び ϕ が出現する文書数 /φ が出現する文書数) である。

[2] キーワード抽出方式については、[1] を参照のこと。

[3] 他に係り受け組を利用する方式 ([2],[3]) もある。

なお、この処理の目的は、共起度の強いキーワード組に絞って、そこから文の内容を推定することである。そのためには、係り受け関係にあるキーワード組や、現象を表すキーワードとその発生原因を表すキーワード組が、強い共起度を持つことが望ましい。そこで、元々の文書をさらに細分化したものをここでの文書単位とし、より狭いスコープでの共起度を算出する。このスコープを決定するためにデータの予備調査を行なったところ、原因と現象に関して、「(原因) のため (現象) が発生した。」などの 1 文内の共起関係が成立する場合が多いが、「(原因) が発生した。そのため (現象) が起こった。」などの表現で 2 文に分けて記述されるパターンも見受けられた。そこで、今回は、文書の各文に分割し (その分割結果を A,B,C,... とする)、1 文ずつずらして 2 文をまとめたもの (つまり、{A,B},{B,C},...) を新たな文書単位として共起度を算出することにした。

次に、共起度の強いキーワード組からグラフを作成する。今回は、キーワード組を上位から順次グラフに追加していき、終了判定は分析者に委ねることにする[4]。

また、特徴キーワードとして上位に出現しているにも関わらず、他のキーワードとの共起度が一律に低いようなキーワードは、この手法ではグラフに追加されずに内容推定が行なわれないという問題がある。これは、そのキーワードと共起するキーワードが一般的なもののみである場合に起こり得るので、漏れのない内容推定には、そのキーワードが出現する文を参照する事も必要となる。今回は、入力したキーワード前後の記述を参照できるツールとして、PortableKiwi[4] を使用した。

3 実験

航空安全レポート 6432 件[5]から、「乗員・乗客の負傷事例」の主な発生パターンを抽出するタスクを実施する。ただし、負傷事例としては乱気流 (Turbulence) を原因とするものはよく知られているので、今回のタスクではそれ以外の負傷事例にフォーカスを当てる。

[4] この自動化は今後の課題である。

[5] 1992 年 4 月から 2003 年 12 月にかけて作成されたデータ。なお、個人のプライバシー保護のため、個人の特定につながる情報は除いてある。

3.1 検索による絞り込み

まず、負傷関連の事例に絞り込みを行なう。この絞り込みのキーワードとして、「負傷, 火傷, 裂傷, 軽傷, 重傷, 切り傷, 挫傷, 創傷, 怪我, 打撲, 骨折」を採用し検索を行なったところ、 234 件の事例が得られた。この 234 件を集合として、特徴キーワードを抽出した所、入力されたキーワードに続いて「Turbulence」「遭遇」「揺れる」といったキーワードが続いており、さらに共起組によるグラフを描くと、これらのキーワードで 1 つのグループを構成しており、確かに乱気流による負傷事例が多いことが確認できた。

次に、この 234 事例から「Turbulence」を含む文書を除き[6]160 件を得た。

3.2 クラスタリング結果からの内容推定

作成した 160 件についてクラスタリングを行なった結果 7 個のクラスタを得た。この中の最大クラスタにおける内容推定について詳細に説明する。

このクラスタの特徴キーワードを求める。その上位 14 位までを表 1に示す。

表 1: 特徴キーワード (上位 14 個)

順位	キーワード	順位	キーワード
1	負傷	8	暴行
2	負う	9	お茶
3	火傷	10	引き渡す
4	殴る	11	サービス終了
5	暴力	12	コーヒー
6	手首	13	重傷
7	男性	14	こぼす

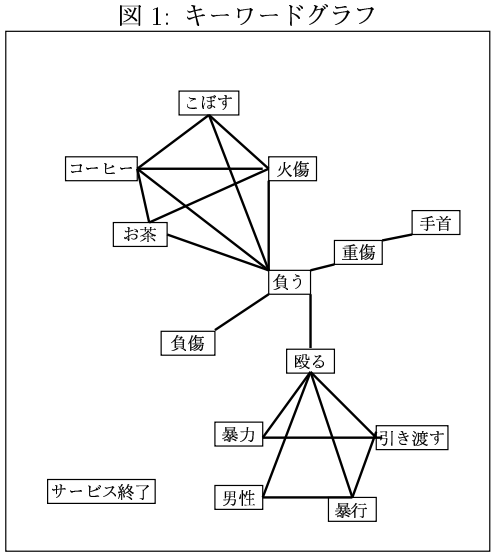
次に、これら上位特徴キーワード間の共起度を計算し、共起度上位 20 位のキーワード組でグラフを描くと図 1を得る。

このグラフより、 2 つのキーワードグループが構成されている事がわかり、各々のキーワードグループから 2 つの発生パターンが推定される。

- 飲物をこぼして火傷

[6] この操作では、「Turbulence というキーワードは含むが Turbulence 自体は発生してない負傷事例」も除かれてしまう。この操作で除かれた 74 事例について内容を精査したところ、 4 事例において Turbulence は発生していなかった。

図 1: キーワードグラフ



- 機内暴力行為による負傷

また、グループに入らなかった「手首」や「サービス終了」については、検索ツールにより元文を参照して発生パターンとなるかを検証する。この例では新規の発生パターンは獲得できなかった。

以上のような操作を各クラスタについて行なう。結果として、例えば、共起関係グラフからキーワードグループとして「歩行 - 貧血 - 転倒 - 裂傷」が得られ、ここから「歩行中に貧血で転倒して負傷」という発生パターンが推定される。また、特徴キーワード「サービスカート」の元文を参照することで、「移動するカートと接触して負傷」という発生パターンが獲得される。

3.3 内容推定結果

以上のような分析操作により推定した内容を、発生パターンとして以下にまとめる。

- 単語間の共起度から推定したパターン
  1. 乱気流を原因とする負傷
  2. 飲物をこぼしたことによる火傷
  3. 機内暴力行為による負傷
  4. 歩行中の貧血による転倒
  5. トイレのドア関連の負傷
  6. 乗員ベッドや手荷物関連の負傷
- 本文を参照して推定したパターン

7 地上走行中の負傷

8 Turbulence 以外の機体姿勢変化を原因とする負傷

9 カートとの接触などによる負傷

4 評価と考察

負傷関連のキーワードを含む 234 件の記述内容を全てチェックし、運航中に実際に負傷者が発生した 136 件が、推定した発生パターンのどれに当てはまるかを検証した (表 2)。

表 2: 抽出パターンの検証結果

パターン	件数	パターン	件数
1	52 件	6	9 件
2	10 件	7	4 件
3	11 件	8	4 件
4	14 件	9	4 件
5	6 件	その他	22 件

パターンに漏れた事例 (その他) が 22 件あり、推定した 9 パターンで負傷事例の 83.8% をカバーしていることになる。また、この 22 件を人手でグルーピングしたところ全て 2 件以下のグループとなり、タスクの目的は達成できたといえる。

さらに、今回の分析対象データに含まれなかった 180 件[7]における実際の負傷事例 5 件に関して、発生パターン 1 が 4 件、発生パターン 2 が 1 件であり、全てカバーできていた。

次に共起関係算出に用いたスコープについて検証する。今回の実験で、強い共起度を持つキーワード組として原因の連鎖が獲得された例としては「貧血 - 飲酒」、「暴力 - 酒」があった。そこで、各々の共起キーワード組を構成する 2 単語を含む文書において、この 2 単語が共起する最小文数を調査したところ、 1 文内共起が 7 件、次いで 2 文:2 件、3 文:2 件、 4 文:1 件、 8 文:1 件となった。この結果からは、スコープは 3 文程度が適切であるといえるが、スコープが広がるにつれ、キーワードグループからの内容推定が難しくなる可能性がある。また、キーワード組により適切なスコープ範囲が異なる可能性も考慮する必要があるので、スコープ範囲に関してはさらに検討が必要である。一方、スコープを広げることで、人間が思いつかないような因果関係が得られる可能性もある。このような因果関係の抽出に関しては、情報抽出技術を用いて原因を表すキーワードを抽出し、その間の共起度を計算することで、より明確な形で因果関係が獲得できる可能性が高い。これらは今後の課題としたい。

また、共起関係に基づいたグラフでなく、より強い関係である係り受け関係に基づいてグラフを作成する方法も考えられる。この際には、係り受け組の統合処理の問題や、複数文に分かれた因果関係の処理などを考慮する必要がある。

[7] 2003 年 12 月から 2004 年 8 月にかけて作成された航空安全レポート。

5 まとめ

多量の文書データから、そこで述べられている主な内容を抽出する分析タスクを支援する方式について提案を行なった。本手法を航空安全レポートとよばれる実データに適用したところ、出現頻度が 3 件以上のパターンは全て検出でき、本手法の有効性を確認できた。

なお、本稿で使用した航空安全レポートは、航空の安全に資するために使用したものである。

謝辞

本研究に対する有用な助言、及び PortableKiwi の使用を許諾いただいた東大・中川先生はじめ中川・田中研究室の皆様に深く感謝いたします。

参考文献

[1] 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

[2] 齊藤孝広: 文書記述内容のメタデータ化について, 情報処理学会研究会報告,FI-70,pp51-58,2003

[3] 内元清貴　他: キーワードからのテキスト生成, 言語処理学会第 8 回年次大会,pp375-378,2002

[4] 藤本宏涼 他: ローカルコーパスからのテキストマイニングツール:PortableKiwi, 言語処理学会第 11 回年次大会,2005