トラブル報告文の事態進展パターンの認識

中田 亨[*1]

Text-Mining on Accidental Reports to Recognize Event Progress

Toru Nakata [*1]

Abstract – To prevent accidents, it is very important to learn why and how past accidents occurred and escalated. The information of accidents is mostly recorded in natural language texts, which is not convenient to analyze the flow of events in the accidents. This paper proposes a method to recognize typical flow of events in a large set of text reports. By focusing two adjacent sentences, our system succeeded to detec typical pairs of predecessor word and successor word. Then we can recognize the typical flows of accidents.

Keywords: text mining, bag of words, natural language processing, and accident prevention.

1. 目的

事故を防ぎ、安全を守る上で、過去の事故事例を学ぶ意義は大きい[1]。事故は、「想定外」や「前代未聞の形」でも起こり得るが、過去に起きた事故の同工異曲の繰り返しが多いことが、産業事故の統計からは見て取れる。また、前代未聞の事故のパターンであっても、部分的には過去の事故事例と共通していることがある。過去の事故事例を多く知っている人は、未知の事故のリスクを予感し、それを防ぐ能力を持てるだろう。

過去の事故事例を学ぶ際には、もっぱら事故報告書や新聞記事といった文章情報に頼らざるを得ない。理想的なのは、できるだけ事故の実態を伝える生々しい未加工なデータから事故の実態を知り、再発防止を考えることである。事故を起こした機械の実物や、事故現場、あるいは動画などが生のデータがあれば考察の題材にしたい。しかし、生のデータは保存や伝達に不向きであるので廃棄されてしまうことが多い。後に残るのは、状況を言葉で記号化し、大幅に要約した文章情報が残されるだけである。

文章情報は、一般には取り扱いに労力がかかるデータ形態である。二つの事故報告書を見比べて、それらの共通点や相違点を特定することは容易ではない。人間がしっかり読解すればその意味内容がわかるものの、そのコストは非常に高い。

反面、世の中には膨大な量の事故報告書や新聞記事が発行され、溜まり続けている。もはや人間が読解し分析できる量を超えており、せっかくの事故情報が活用されぬまま溜まっていくだけである。

もしこれが､コンピュータを使って高速に読解を行い、共通点や相違点、事故の因果関係を認識し、事故の予防のための教訓を出すことができれば、安全の増進に大いに役立つであろう。こ

現在、発展が著しい自然言語処理技術を使えば、事故情報の自動分析は可能ではないかと思われる。この一手法を提案することが本稿の目的である。

2. 提案技法

2.1 因果関係分析の難しさ

事故とは、因果関係を持った一連の出来事の流れであり､その結果が人間にとって不本意なもののことを言う。

因果関係を知ることが、事故を分析し再発防止に最重要である。悪い出来事を防ぐには、その原因となった出来事の発生を防げばよいが、これは因果関係が分からなければできない。また、異なる事故事例の共通点や相違点を識別し、事故のパターンを分類するためにも、因果関係を重点的に分析することが重要である。

このため安全工学では、出来事の因果関係を、グラフ化して表現する技法が、自己分析の基本とされる。Event Tree Analysis (ETA), Fault Tree Analysis (FTA), HAZOP といった代表的な事故分析・リスク分析技法は、いずれも出来事の因果関係をグラフ化することを分析作業の基盤としている[2]。

これらのグラフを作成する作業は、安全工学の専門家による慎重な検討を必要とし、彼らの手作業によってなされているため、非常に手間がかかり、処理できる件数にも限りがある。例えば、産業技術総合研究所のリレーショナル化学災害データベース(RISCAD)という事故情報データベースでは、化学産業での事故の記録を公開している。データの総数は 8,438 件(随時追加されつつある)であるが、そのうち出来事の流れをグラフ化した「事故進展フロー図」まで作成されているのは 162 件しかなく、全体の 2%である。大半はテキストのみの記述であり、因果関係は読者がテキストを読解して判断するしかない。

このように、手動によって事故報告文から因果関係の

*1: 産業技術総合研究所　人工知能研究センター
*1: Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST).

グラフを作成することははかどっていない。ここで、自動化することを考えてみよう。

自然言語処理技術では、イベント認識の研究が盛んである。文章中から、そこに書かれている出来事を識別し、その因果関係等を知る技術が盛んに研究されているが、その困難さも明らかになっている。

そもそも因果関係の定義は難しい[3]。「青酸カリを飲むと死ぬ」という因果関係は誰もが認めるだろう。だが「クレオパトラが生まれていなければ、坂本龍馬は生まれなかっただろう」という関係は、正しいそうではあるものの、通常我々はこれを因果関係とは見なさない。また、「A さんが 1 歳年を取ると、B さんの年齢も 1 つ上がる」といった偽相関は、単に相関係数を見ているだけでは排除できず、内容に踏み込んだ深い考察が必要になる。いわゆる因果関係を、我々の常識に合うように、しかも統計学的な観点からも明確に定義することは容易ではない。

2.2 出来事の計数の困難

出来事の個数の計数の仕方も一意ではない。自然言語処理で用いられる素朴な方法は、「動詞 1 個を 1 つの出来事」や「1 文を 1 つの出来事」としてカウントすることである。「大根を洗る。次に切る。そして煮る」という表現では文が 3 個あるから、3 つの出来事があると見なすという流儀である。しかし、「洗浄され切断された大根を煮る」とすると、見かけ上、文は 1 個に減ってしまう。また「大根を料理する」という包括的な表現を使うこともある。この場合、執筆者の認識では、細かい一挙手一投足の動作には言及せず、「料理する」という包括的な現象のレベルで認識しているのである。このように執筆者によって「出来事の粒度」が変化する。

事故を説明するのに適切な粒度を、客観的に定めることは難しい。

2.3 簡便化の方法

因果関係の厳密な定義はあきらめ、何らかの用途には利用できる程度で簡便に扱うことを考える。「まず A が起きた。次に B が起きた」という報告の件数が際立って多ければ、「A の後には、B が起こりやすい」という関係があると、見なすのである。

さらに本稿では、報告書文中において単語 A が登場した後、あまり間を置かず、単語 B が登場することが多ければ、「単語 A を含む出来事の後は、単語 B を含む出来事が起こりやすい」という関係があると判定する。

かなり簡便な方法であるから、当然、本来は因果関係が無いペアを誤検出する。しかし、サンプル数が大きくなれば、大数の法則によって、偶発的な誤検出は淘汰され、真の因果関係が多く残ると期待できる。

本稿では、「執筆者は、1 つの文は 1 つの出来事を表記する。各文では適切な粒度で出来事が記述されている」と仮定するにとどめる。仮定の妥当性は、実験によって検証する。

2.4 隣接２文カップリング

本稿で提案する技法は、事故報告書の文章中にて隣り合う 2 つの文のペアを分析にスコープとし、その中での単語の共起性を計測することで、出来事の流れ・因果関係を洗い出すものである。

この技法は、事故を報告する文章が次のような特徴を持つことを仮定している。

- 1 つの文は「1 個の出来事」が書かれている。
- 文章は出来事を時系列に記述している。すなわち、文の順序は、それらが表す出来事の発生順序と同じである。

全ての事故報告書が、この特徴を満たすわけではないが、この性質を備えているものは多い。特に、事故の目撃者や当事者の陳述をいくつか読んでみると、この特徴を備えている事例が大半であることに気がつく。

さて、ある事故を報告する文章がN個の文から成り立っているとする。今、i番目の文S_iと、$i+1$番目の文S_{i+1}に注目する。S_iとS_{i+1}との範囲で bag-of-word (BoW)を考え、単語の共起を観測する。図 1 の赤カッコのように、隣接する 2 文ごとにスコープを定めて計測するのである。

すなわち、この 2 文の中に出現する単語を調べ、同じ BoW 内に出現した単語のペアをカウントする。カウントする単語の品詞は、名詞、固有名詞、動詞、形容詞に限る。

例えば、「液を排出した。タンクが爆発した。」という一節が文章にあれば、(液、排出)、(液、タンク)、(液、爆発)、(排出、タンク)、(排出、爆発)、(タンク、爆発)という単語ペアの共起頻度を 1 つずつ増やすのである。

また、単語共起が報告書内でどこに出現するか、その位置の情報を分析に加味する工夫も施す。報告文の第 1 文の前に第 0 文を置き、最終文S_Nの後にS_{N+1}を置く。第 0 文は「StoryStart」というマークのみ、S_{N+1}は「StoryEnd」というマークのみからなる文である。StoryStart と多く共起する単語は、報告文の第 1 文に多く出現する傾向がある。さらに伝播を考えることができて、第 1 文に出現する傾向がある単語に対して頻繁に共起する単語は、報告書の第 2 文に出現する傾向がある。そして、第 2 文での頻出語と共起するのは、第 3 文に頻出する語である。このように、各単語の頻出出現位置が示唆されることになる。

複数の事故報告文に対して、以上に説明した方法で共起頻度を計測する。こうすると、異なる事故でも共通して登場する共起に組み合わせが、より多くカウントされる。よって、事故事例の間に共通して登場しやすい単語の前後関係が浮かび上がる。

この共起頻度の分析結果を、多次元尺度法などを用いて可視化すれば、事故報告文の冒頭から末尾にいたる過程で、どの前後の順番でどのような単語が登場しがちで

あるかが明らかになる。

なお、本手法では、各 BoW がどの報告文に属するかという情報は無視する。単に隣接 2 文からなる BoW の集合に対して分析を実施するだけである。特定の事故事例の開始から終末までを通しで、話の流れを追跡し、読解するわけではない。BoW は図 1 にように近隣の BoW 同士でオーバーラップを持っているとはいえ、特定の事例の話の流れを再構成することは難しくなる。

これは、特定の事例の話を流れだけに分析が捕らわれる作用を緩和し、様々な事故事例の中で最も頻発する、単語の前後関係を抽出する効果が期待できる。例えば、「A をした。B が起きた」という BoW と、「B が起きた。C が起きた」という BoW が数多く存在した場合、A → B → C という出来事の流れが存在すると出力する。たとえ、この 3 組を含む事故事例が元のデータ中に存在しなくても、起こりえる出来事の流れとして出力するのである。こうして、まだ発生していない事故のパターンであっても、そのリスクを検知できるだろう。

3. 実験

3.1 事故報告書データ

今回、事故報告文として採用したのは、航空におけるヒヤリハットデータを収集している、米国航空安全報告制度(Aviation Safety Reporting System, ASRS)のデータである。2013 年に発生した全てのインシデント（軽度な被害で済んだ非正常な現象）に関する報告文を用いた。このデータには、パイロットなどのインシデントの当事者による陳述(narrative)という部分がある。これは当事者の一人称で、事故の経過を叙述しているものである。図 1 に陳述データの一例の冒頭部を掲げる。

実験に用いた陳述データの概要は次の通りである。

- 言語: 英語
- 書き手 : インシデントの当事者である航空関係者 (パイロット、管制官、整備係等)
- 報告文件数: 4,469 件
- 単語数: 1,365,260 語。うち分析に使用したのは 823,722 語。
- 異なり語数: 28,615 語。うち分析に使用したのは 21,614 語。
- 文の数: 110,963 本。
- 1 つの報告は平均で、305 単語、25 文からなる。

なお文の区切りについては次のような工夫をほどこした。通常はピリオドの出現をもって文を区切るが、今回はセミコロンも文の区切り文字として採用とした。これは、一連の動作をセミコロンを使って区切った複文で表現している事例が多かったためである。

陳述では、書き方の統一性や、詳細度、報告の長短は、必ずしも統一されているわけではないが、報告者は航空分野に詳しいパイロット等の資格者がほとんどであり、ASRS の本部による内容チェックはあるので、品質に一定の安定性は存在すると言える。

なお、ASRS では、陳述データを因果関係のグラフ化に変換して発表することはほとんどしていない。陳述データはそのままの形でもインシデントを生々しく伝え、読者をドキリとさせ、リスクを知らしめる効果がある。よって、知っておくべき事故パターンについての代表的な陳述を、メールマガジン等で航空関係者に配信している。

- **<<StoryStart>>**
- **A gate agent Supervisor entered the aircraft 11 minutes before the scheduled departure time and wanted to close the aircraft door.**
- **I explained that overhead bins were still open as the last of the passengers had just entered.**
- **And a long line in the aisle was still present.**
- **I told her that it was 11 minutes until scheduled departure.**
- **I had also informed the gate agents twice during boarding that oversized carry on baggage was entering the aircraft.**
- **The agent Supervisor stated that the cargo bin had already closed.**
- **I told her that we may still have bags to check due to being oversized.**
- **At that time a bag came to the forward galley that was obviously well larger than the bag sizer would allow.**

図 1　事故報告文データ事例と、隣接 2 文の BoW
Fig.1 An example of narrative report of ASRS. Red braces indicate Bow of two adjacent sentences.

3.2 分析の手法と結果

形態素解析、単語の抽出、共起頻度の計測、多次元尺度構成法の作図に KH Coder [4]を用いた。

まず、比較のため従来手法として典型的な、各報告文を BoW として計測した単語共起の観測を試みた。

その結果が図 2 である。共起しやすい単語のクラスターを 7 個抽出した。(7 はクラスター分析でしばしば選ばれる分類の数であるので採用したが、特に意味は無い。) クラスターそれぞれの意味は、その内容から判断すると、次のように整理できる。

1. 保守作業と乗客 (単語例: passenger, maintenance, takeoff)
2. 地上移動 (taxi, ground position)
3. 他の飛行機・通信 (traffic, radio, frequency, cross)
4. 緊急事態 (emergency, problem, declare, decide, Quick Reference Handbook (QRH))
5. 対処 (procedure, incident, receive, action)

6. 自機の操縦 (course, climb, level, knot)」
7. 動詞全般 (take, use, get, call, make)

通常、こうしたクラスター分析は、テキストを分類するために行われる。すなわち、各報告文に登場する単語群は、1つないし少数個のクラスターの単語群に偏っていることが多い。そうでなければクラスターが明瞭には形成されないからである。よって、各テキストが、いずれかのクラスターの単語群を多く含むかによって、その主題を識別して、他のテキストと分類できると期待される。

今回の分析結果を見てみると、確かに「地上移動」中のトラブルの話や、「他の飛行機・通信」に関するトラブルの話というように、報告文を分類することは妥当と思われる。従来からもよく行われてきた文書分類法の一つである。反面、この分析結果は頻出単語が識別できる程度に留まっており、事故のストーリー展開が分かるほどには深いとは言えない。

次に提案手法での結果を見てみよう。図3に結果を掲示する。上述した従来手法と同様の処理であり、BoW が隣接する 2 文という点だけが異なる。

7つのクラスターに分けた結果、それぞれに次のような意味が解釈できた。

1. 離陸や高度といった飛行の状況、あるいは天候を説明する単語群。 (単語例: takeoff, level, climb, weather. begin, start) このクラスターは報告文開始のマークを含んでいたので、報告文の冒頭に多い単語の集合といえる。
2. エンジン、車輪、緊急事態等に関する単語群。 (engine, gear, emergency, declare, QRH, return)
3. 空港での事物に関する単語群 (airport, runway, Air traffic control (ATC))
4. 飛行コースの設定・変更に関する単語群 (degree, clearance, cross, hold, change)
5. 他機や、通信に関する単語群 (traffic, controller, tower, frequency, call, hear, ask)
6. 思い込みや知識、要望など、思考に関する単語群 (believe, think, feel, need)
7. インシデント報告、保守作業、人員に関する単語群。 (incident, problem, issue, inform, maintenance, crew, passenger) このクラスターは報告文終了のマークを含んでいたので、報告文の末尾に多い単語の集合といえる。

報告文の話の流れは、図中の報告開始マークから報告終了マークに向けて、すなわち第 1 クラスターから第 7 クラスターに向けて進む傾向が、必ず全てで成り立つわけではないが、分析の上では最も頻発するパターンであると言える。

話の流れで最も典型的とある連結関係を図 4 に示す。大まかに次の 3 つのパターンが存在すると解釈した。

1. 「離陸許可、出力表示」→「地上移動、燃料、機長、副操縦士、緊急参照マニュアル、着陸装置、エンジン、緊急事態、宣言、引き返し」→「空港のゲート、保守修理」
   【話の流れ】エンジンや着陸装置のトラブルにより、空港に引き返し、保守作業要員を呼んだ。
2. 「開始、停止、降下」→「滑走路、管制塔、接近、空港」→「問題、通知」
   【話の流れ】空港からの離陸、あるいは空港への接近時に、滑走路で他機との進路干渉が起こり、インシデントとして報告した。
3. 「上昇、オートパイロット、ノット」→「コース、マイル、間隔、変更」→「他機、管制官、周波数、呼びかけ」→「思う、信じる、知る、感じる」→「問題、状況、気付く」
   【話の流れ】巡航中に、他機が近づいてきたので呼びかけたが、通信電波の周波数が合わないのか通じておらず、大丈夫と思い込んで飛行を続けていたら、異常に接近してしまった」

いずれのパターンも頻発する典型的なインシデントのストーリーであり、この分析結果は現実と矛盾しないと考える。

従来手法として紹介した単語の共起性の分析だけでは、出来事の時間順序性に関する情報が抜け落ちていた。提案手法では、隣接 2 文しか収めていない BoW であるにもかかわらず、時間順序性の情報が反映されており、ストーリーを推定することができた。

この分析は、各単語がインシデント報告文の序盤・中盤・終盤のどの位置に出現しやすいかという特性を計測したことに他ならない。本来、単語は位置とは関係なく出現してもよさそうである。特に、頻用語は、報告文のどこにでも出現するはずである。しかし、計測したところ時間順序性が検出できた。各単語について、その直前の文に登場しやすい単語や、その直後の文に登場しやすい単語の傾向が存在することがわかった。

こうした性質が存在し検出できるか否かは、2.2 節で述べた事故報告文の形態と品質次第であって、事故報告文の書かれ方に依存する問題となる。今回は意味のある結果が得られていることに鑑みると、ASRS の報告文の書き手は概ね、時間前後性の情報を崩さないように、1 つの出来事を 1 つの文にまとめ、それらを時系列に配列していたと認められる。

隣接する 2 文での BoW という単純かつ局所的な観察の積み重ねだけで、話の流れという大域的な構造が出現したことは意外であった。これは、事故報告書のサンプル数が多く、かつ特定の狭い産業分野に絞ったため、頻出する話の流れのパターンが雑音に埋もれることなく検出できたと考えられる。

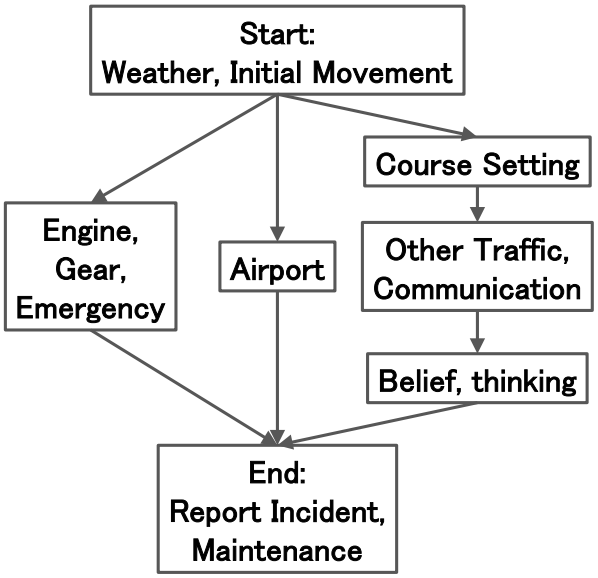

図 4　事故報告文の「出来事の流れ」の傾向

Fig.4 Typical pattern of “flow of events” found in aviation incident reports.

4.　結論

4.1　結果のまとめ

本稿は、事故の防止の目的で、人手では読み切れないほどの大量の事故報告文の自動的処理し内容理解する方法を提案した。それは、事故報告文において隣接する２つの文に注目し、その中での単語間の共起頻度を計測し、多次元尺度構成法でグラフ化することで、事故全体での話の流れを検出するものである。

航空インシデントの報告 4,468 件を分析の結果、典型的な話の流れを検出することができた。

このように、膨大な量の報告文であっても、自動的に分析し、産業の現場において頻発しがちな事故進展のパターンを抽出することができた。

4.2　将来課題

本稿では処理が煩雑になるので採用しなかったが、より精密に分析するに、S_iに出現した単語Aと、S_{i+1}に出現した単語Bのペアを考え、AからBへの順序性情報を備えたペアをカウントすることは、さらに優れた分析技法になると期待している。この処理によって、ある単語が出現した文の次の文で、どの単語が出現しやすいかという遷移行列が得られる。この遷移行列を特異値分解すれば、頻度の高い単語のつながりの筋が検出でき、それは頻出する事故パターンを意味するだろう。

ただし、この行列の行や列のサイズは報告文中でカウントする語彙の総種数になり、一般には非常に巨大である。このような巨大行列の特異値分解は、最近の計算機の能力を以てすれば不可能ではないが、やはり容易にできるものではない。

各産業分野に応じた適切な語彙の絞り込みと、オントロジーの整備による同意語や含意関係の解決が、精度向上と計算の高速化に必要になるだろう。

謝辞

この成果は、国立研究開発法人新エネルギー・産業技術総合開発機構(NEDO)の委託業務の結果得られたものです。

参考文献

[1] 化学工学会 SCE Net 安全研究会, CCPS: 事例に学ぶ 化学プロセス安全 Beacon の教訓と事故防止の知恵, 丸善出版, (2015).

[2] International Electrotechnical Commission: IEC 61508-7:2000 Functional Safety of Electrical Electronic/ Programmable Electronic Safety-Related Systems — Part 7, Overview of Techniques and Measures, (2000).

[3] Judea Pearl: Causality: Models, Reasoning and Inference, Cambridge University Press, (2000).

[4] 樋口耕一: 社会調査のための計量テキスト分析, ナカニシヤ出版, (2014).

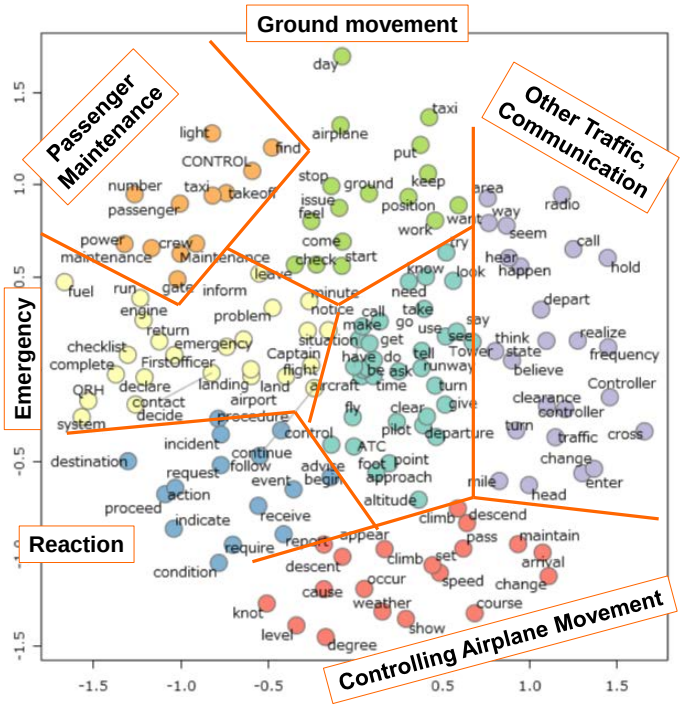

図２　事故報告文を BoW データとした場合の単語共起性多次元尺度構成法とクラスタリング結果

Fig.2 Conventional analysis result of multi-dimensional scaling of co-occurrence of words in each report. Colors and labels indicate their clustering.

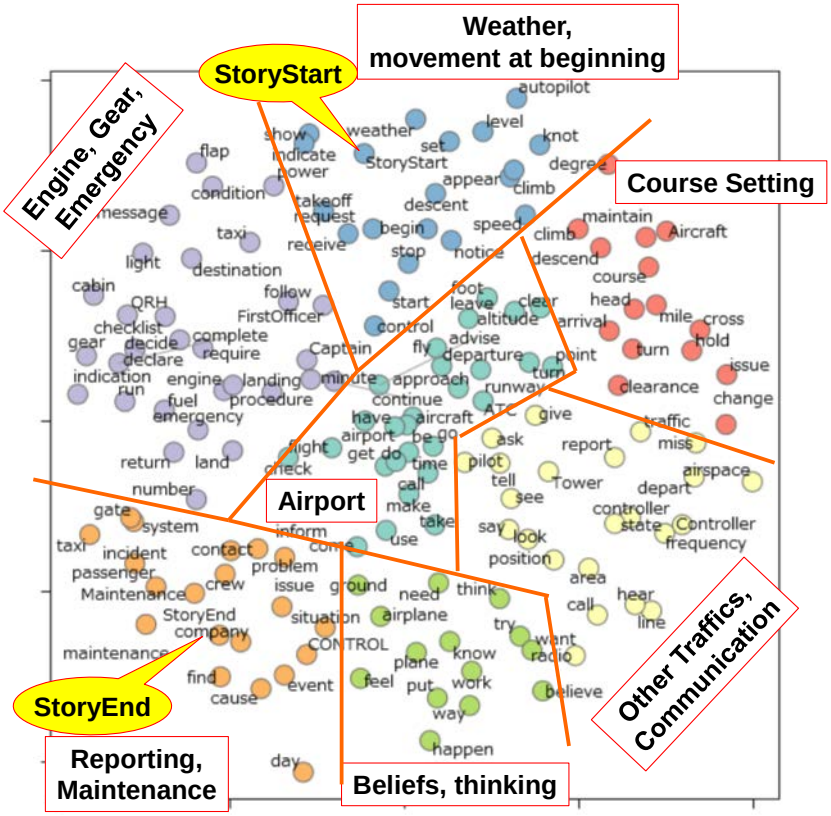

図３　隣接 2 文を BoW データとした場合の単語共起性多次元尺度構成法とクラスタリング結果。StoryStart が報告文の文頭、StoryEnd が文末である。

Fig.3 Result of proposed method. “StoryStart” stands for the beginning of each report, and “StoryEnd” is for the end.