日報を対象とした障害予知

柿元 芳文 , 山本 和英
長岡技術科学大学 電気系
E-mail:{kakimoto,ykaz}@nlp.nagaokaut.ac.jp

1　はじめに

多くの企業では、社員の勤務状況を勤務日報という形で報告させ、管理している。近年では文書の電子化が進み、勤務日報を Web や E-mail で報告させる企業が増えている。

日誌を閲覧する際、直接的に報告されている障害に対しての対策は容易であるが隠れた障害、次に起こり得る障害を人間が発見し対策することは極めて困難である。隠れた障害を見落とし、対応が遅れてしまうことによって企業の信用を失う事態に発展してしまう可能性がある。もし次に起こりうる障害を予知し、注意を喚起できるならば日報の閲覧コストを下げると共に障害が表面化する前に対処することが可能となる。我々は日報を入力として、次に起こり得る障害を予知するシステムを提案する。

2　関連研究

障害を表す表現を抽出する研究として、De Seager ら [1] は統計的情報を用いて障害を表す名詞（トラブル表現）の収集を行っている。「トラブル」「災難」などの明確な障害を表す単語や、否定形の動詞との共起情報を素性とし、SVM に学習させることでトラブル表現を抽出している。さらに、得られたトラブル表現と共起しやすい名詞を結びつけることで「薬-副作用」のような object-trouble ペアを収集している。しかし、このペアは名詞の対である。よってある日報からこのペアが抽出出来たとしても、そのトラブルが実際に発生しているかどうかは判断できない。本研究では、障害の発生を示し障害の対象と状態を内包している表現を獲得し障害情報辞書を構築した。

自然言語処理の分野では障害を予知し提示する研究は、現在行われていない。しかし、二つの出来事間の因果関係を同定する研究は行われている [2] [3]。因果関係とは、二つの出来事が原因と結果の関係にあるものを指す。例を例 1 に示す。

例 1) 因果関係の例
入力文：私は、熱が出たため病院へ行った。
因果関係：熱が出る (原因)　病院へ行く (結果)

例 1 では、「病院へ行く」という出来事を「熱が出る」という出来事が引き起こしたと考えられる。これらの手法を用いて障害を表す表現間の因果関係を同定すれば障害の予知も可能となる可能性がある。しかし、因果関係は原因と結果の対であるので、人間にとっては想像が容易な場合が多い。障害の予知を行う場合、人間が容易に想像できる事柄を予知しても意味がない。つまり、人間が気付きにくい障害を気付かせることが、予知システムの役割であると考える。よって、本研究では因果関係は使わず、文書分類の手法を使って予知を行う。文書分類の手法を用いることにより、因果関係で得られるような人間が容易に想像できる障害だけではなく、人間が気付きにくい障害も予知することができる。

3　用いたデータ

本研究で対象としているデータは、企業の日報である。しかし、企業の日報には企業機密や個人情報などが含まれており、研究対象として入手のは難しい。また、研究に用いるためには、大規模なデータが必要である。これらの理由から、本研究では Web 上のデータを日報として用いている。

本研究では価格比較サイトである「価格.com[3)]」のデータを日報とみなして用いる。よって、本論文中で「日報」とは「価格.com」の一記事を示す。価格.com の全ての書き込みには、書き手が設定するタグが存在する。これらのタグは人手でつけられているため、信頼できる情報である。本研究では、5 節の学習データの作成時にこのタグを利用している。

4　手法概要

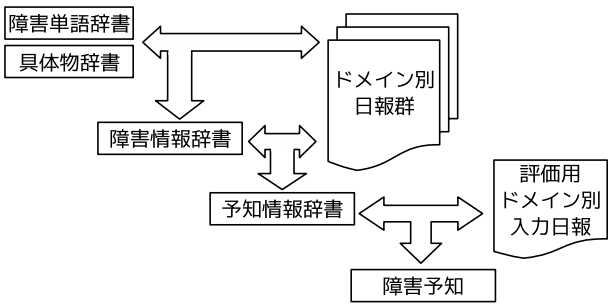

図 1: 提案手法の処理の流れ

本手法は、ある企業の日報が複数のドメインについての情報を含むことはないと考える。よって、本手法ではドメイン別の処理を行っている。

本手法は大きく分けて三段階からなる。まず、予知すべき障害を障害情報と定義して学習データから抽出し、障害情報辞書を構築する。次に障害情報辞書と学習データを基に、障害情報にその障害が起きた時点の状況を付与し予知情報を作成する。各障害情報の予知情報を集めて予知情報辞書を作成する。そして、入力日報と各予知情報との間で類似度の計算を行い、入力日報と障害情報の対応付けを行う。最後に、この対応付け手法を基にして入力日報に起こり得る障害を予知し、出力する。

5　障害情報辞書の構築

5.1　障害情報の定義

本手法では、障害情報を問題解決への対処に役立てようと考えている。よって、障害情報は生じた障害の内容を推察することが可能な単位でなくてはならない。

障害が発生したことを表す単語は多く存在する。例として、「壊れる」や「遅延」などが挙げられる。これらの単語を手がかりとすれば、何らかの障害が起こったことを推察することが出来る。しかし、「椅子が壊れる」と「サーバーが壊れる」では危険度も取るべき対応も大きく異なる。つまり単語は障害の内容を推察する表現として不十分である。障害の内容を推察するには、障害が起こっている対象とその対象の状態が必要だと考える。そこで、本研究では障害情報の単位として構文片 [5] を用いる。構文片とは構文解析結果の修飾要素と被修飾要素の対を基にしたものである。

構文片は二文節からなり、障害の対象と状態を含むことが出来る。しかし、全ての構文片が障害の対象と状態の対を持っているわけではない。よって構文片に以下の構文パタンを設け、このパタンに合致する構文片を障害情報とした。

具体物名 + 格助詞「が」 ⇒ 障害を示す単語を含む文節

このパタンにより、前項で障害の対象、後項で障害の状態を表す構文片を障害情報として集めることが出来る。

5.2　障害情報抽出手法

障害情報抽出手法の流れを図 2 に示す。

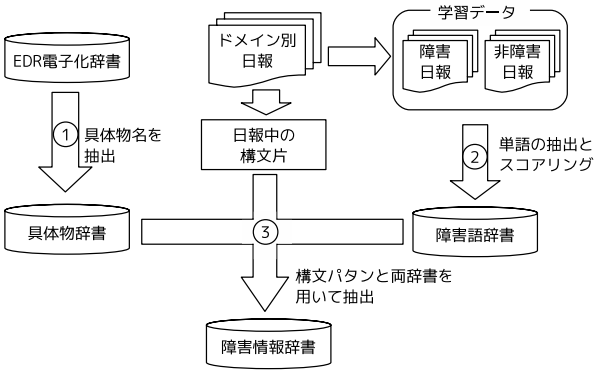

図 2: 障害情報抽出手法

本手法で用いる学習データをドメイン別日報から構築した。障害日報とは、価格.com で「悪い」タグが付けられた日報であり、非障害日報とは「良い」タグが付けられた日報である。

具体物辞書は障害情報の「具体物名」の同定に用いる。具体物辞書は EDR 電子化辞書 [4] を基に構築した。EDR 電子化辞書が持つ意味のカテゴリを人手で観察し、障害情報の「具体物名」に相応しいカテゴリを選別した。選別の基準は「機械や道具などを表す、又はその一部分を示す単語」とした。

障害語辞書は、障害情報の「障害を示す単語を含む文節」の同定に用いる。障害語辞書は学習データ中の内容語[1]に障害情報らしさのスコア S_{tc} を付与することによって構築した。スコアの算出は、藤村ら [6] の手法を参考とした。藤村らの手法を本研究に適用した場合の式を式 (1) に示す。

$$S_t(w_i) = \frac{P'(w_i) - N(w_i)}{P'(w_i) + N(w_i)} \tag{1}$$

$$P'(w_i) = \frac{P(w_i)}{P_{doc}} \times N_{doc} \tag{2}$$

w_i はある単語を表す。$P(w_i)$ はある単語が出現した非障害日報の数を示す。$N(w_i)$ はある単語が出現した障害日報の数を示す。P_{doc},N_{doc} は非障害日報の総数、障害日報の総数をそれぞれ表す。式 (1) により算出したスコアが負の場合、その単語は障害日報に出現しやすい単語である。しかし、式 (1) では単語の学習データ中での出現頻度による差異が反映されていない。よって、出現頻度による差異を考慮するため、確率の信頼区間推定法 [7] を用いた。信頼区間を考慮した式を式 (3) に示す。

$$S_{tc}(w_i) = \begin{cases} S'_t(w_i) - S_c & S'_t(w_i) \geq 0 \\ S'_t(w_i) + S_c & S'_t(w_i) < 0 \end{cases} \tag{3}$$

$$S_c = 2 * 1.96 \sqrt{\frac{S'_t(w_i)(1 - S'_t(w_i))}{P'(w_i) + N(w_i) + 4}} \tag{4}$$

$$S'_t(w_i) = \frac{P'(w_i) + 2}{P'(w_i) + N(w_i) + 4} - \frac{N(w_i) + 2}{P'(w_i) + N(w_i) + 4} \tag{5}$$

以上より学習データ中の単語にスコア S_{tc} を付与し、スコアと共に障害語辞書に登録した。ただし、信頼区間を考慮することでスコアの極性が反転した単語については、辞書に登録しない。また本手法では、障害語辞書内の単語に否定語「ない」を考慮する。同じ文節内で否定語「ない」と共に出現した場合、スコアを反転する処理を行う。

上記の具体物辞書と障害語辞書、構文パタンを用いて障害情報を抽出し、障害情報辞書を構築する。ドメイン別日報から構文片を抽出し、前項が具体物辞書内の単語を含む構文片を得る。次に後項に障害語辞書内の単語を含み、そのスコアが負となる構文片を得る。得られた構文片の前項の格助詞が「が」であった場合、障害情報として収集し、障害情報辞書に登録した。

[1]形態素解析器「茶筌 [1]」において名詞, 動詞, 形容詞, 副詞, 未知語となった単語とする。

6 日報と障害情報の対応付け

本節では、5 節で抽出した障害情報を入力日報と対応付ける手法を述べる。ある障害を他のテキストデータに対応付けるという研究は、現在行われていない。しかし、ある文書を他の文書または文書群と対応付ける文書分類という研究は多く行われている [8] [9]。本研究では日報と障害情報の対応付けを文書分類のタスクとして処理する。処理の流れを図 3 に示す。

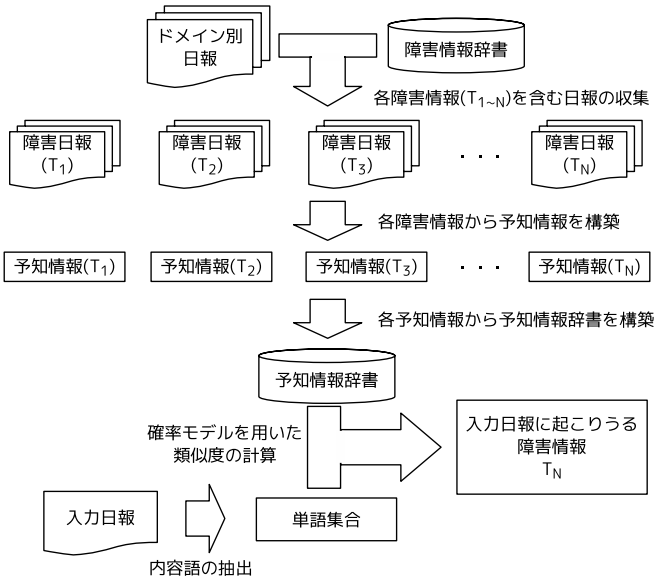

図 3: 障害情報対応付け手法

6.1 予知情報辞書の作成

本節では、障害情報辞書を用いた予知情報辞書の作成手法を述べる。予知情報辞書はドメインごとに構築する。予知情報は、各障害情報に対して付与する。学習データとして各障害情報を含む日報を収集し分類した。本研究では、予知情報の素性として内容語を用いている。各障害情報がもつ日報群から内容語を取り出し、その内容語の頻度と共に予知情報の要素として用いた。ただし、障害情報を構成している内容語は要素から除いた。得られた予知情報を予知情報辞書へ登録した。予知情報の例を例 2 に示す。例 2 は「液晶が ⇒ 割れる」が障害情報であり、「落とす, 指, 衝撃, ポケット, 当たる」が障害情報と共起した内容語である。また、各内容語に付与している数字は各内容語の頻度 (延べ数) である。

例 2) 予知情報の例

液晶が ⇒ 割れる { 落とす.6 , 指.6 , 衝撃.2 , ポケット.2 , 当たる.2}
カーナビが ⇒ 壊れる { 走行.5 , 取り付け.5 , 安い.2 , すれる.1 , 太陽.1}

6.2 日報と障害情報の対応付け手法

本節では、予知情報辞書内の予知情報と入力日報の対応付け手法について記述する。対応付けは、予知情報と入力日報の内容語集合間に対して類似度を算出することで実現する。本手法では、対応付けには以下の二点を重要視する必要があると考えた。

1. 予知情報と入力日報の内容語の一致率
2. 各ドメインでの障害情報の出現しやすさ

予知情報と入力日報の内容語の一致率は、両集合内で用いられている内容語が一致すればするほど類似性が高いという考えから成り立っている。各ドメインでの障害情報の出現しやすさを考慮する理由は、各ドメインにはそれぞれ「起こりやすい障害」が存在するためである。例えば、電話ドメインでは「電源が ⇒ 切れる」カメラドメインでは「シャッターが ⇒ 壊れる」が挙げられる。学習データ内で多く出現した障害情報は、予知情報の要素数も巨大となる。よって、本手法では要素数が大きい予知情報を「出現しやすさが高い」として扱っている。以上のことを考慮し、

予知情報と入力日報間の類似度を測る独自の確率モデルを作成した。計算式を式 (6) に示す。

$$P(f|d) = P'(f|d) - 1.96\sqrt{\frac{P'(f|d) \times (1 - P'(f|d))}{|F| \times |D| + 4}} \quad (6)$$

$$P'(f|d) = \frac{|W_F| \times |W_D| + 2}{|F| \times |W| + 4} \quad (7)$$

f はある予知情報、d は入力日報を示し、F、D はそれぞれの要素の集合を示す。W_F は集合 F に含まれる要素中、集合 D にも含まれている要素の集合である。W_D は集合 D に含まれる要素中、集合 F にも含まれている要素の集合である。

式 (6) の第一項で予知情報と入力日報の内容語の一致率を数値化している。第一項は「予知情報と入力日報から一要素取り出した時、それがどちらも共通の要素である確率」を意味する。また、第二項で各ドメインでの障害情報の出現しやすさを算出している。第二項は 5.2 節で用いた信頼区間推定法の式と同じである。信頼区間は、確率算出の際の試行回数が多い程狭くなる。つまり、式 (6) では $|F| \times |W|$ の値が大きい程信頼区間が狭くなる。$|F|$ はある予知情報がもつ要素集合であり、これはある障害情報が出現しやすいほど大きな集合となる。よって、間接的にではあるが、各ドメインでの障害情報の出現しやすさを数値化していると考えることができる。また、第二項は「出現しにくい障害情報に対するペナルティ」という位置づけであるので、負方向の信頼区間のみを考慮している。

式 (6) を用いて入力日報と予知情報の類似度を計算し、上位 3 件を入力日報から起こり得る障害として対応付ける。

7 障害予知手法

本節では、入力日報から起こり得る障害を予知する手法を記す。本手法は二段階の処理からなる。まず、入力日報が予知の必要な日報であるかを判断する。そして、予知が必要と判断された場合、6 節に示した手法を用いて障害情報を対応付け、予知として出力する。よって、本節では入力日報が予知の必要な日報であるかを判定する手法のみを記述する。

7.1 要予知日報判定

入力される日報は、その全てに予知が必要であるわけではない。日報の中には、障害を導くような日報もあれば良い状況を報告しているような、障害とは結び付かない日報も存在する。6 節に示した手法は、全ての入力日報に対して障害情報を対応付けてしまう。よって、対応付ける前に入力日報が予知の必要な日報かどうかを判断する必要がある。本手法では、予知の必要な日報を要予知日報と呼ぶ。

障害は、良い状況から誘発されるとは考えにくく、やはり悪い状況から起こりやすいと考える。言い替えれば、要予知日報では、悪い状況で使われやすい単語が多く含まれ、障害が起こらない日報では、悪い状況で使われやすい単語が含まれにくいはずである。

この仮定を基に、要予知日報判定手法を構築した。要予知日報判定に用いる数式を式 (8) に示す。

$$S_p = \sum_i S_{tc}(w_i \in D) \quad (8)$$

w_i は障害語辞書に含まれているある単語を示す。D は入力日報内の単語集合を示す。$S_{tc}(w_i \in D)$ は入力日報に含まれていて、かつ障害語辞書にも含まれている単語の障害語らしさのスコアを示す。このスコアは式 (3) に示したものと同一である。

式 (8) は、入力日報内の負のスコアの単語割合が大きい程、S_p が負のスコアとなりやすい。本研究では、式 (8) を用いて入力日報にスコア付けし、S_p が負となった日報を要予知日報だと判定した。

8 評価実験

8.1 評価実験 1 : 人間による障害予知

本評価実験では、入力日報に対する障害予知を人手で行った。評価者は、評価用の日報である 200 件の日報を読み、以下の選択肢の中から一つを選ぶ。評価用の日報 200 件はドメイン「電話」「カメラ」「車」「ゲーム」から 50 件ずつ収集したものである。

- 障害が予知出来る
- 障害が起こり得るが予知は出来ない
- 障害は起こり得ない

また、障害が予知出来る場合は、予知できる障害を最大三つまで書くことが出来る。この実験により、要予知日報判定部の評価を行うことが出来る。また 入力日報に対する「人間でも可能な予知」を収集することが出来る。

8.2 評価実験 2 : システムによる障害予知の評価

本評価実験では、システムが出力した予知がどの程度正しいのかを評価することを目的として行った。本評価実験では、入力日報とともに以下の選択肢が表示される。

(1) システムの出力した予知 3 件
(2) ランダムで出力した予知 3 件
(3) 障害は起こり得るが出力の中にはない
(4) 障害は起こり得ない

(2) のランダムとは、各ドメインが保持している障害情報辞書から、障害情報をランダムで 3 件選択し、出力した場合を示す。評価者は、システム又はランダムの出力に起こり得る障害が含まれていた場合、選択する。この場合は複数選択が可能である。システム又はランダムの出力に起こり得る障害が含まれていなかった場合は、評価者は (3) もしくは (4) を選択する。

本評価実験を用いて、障害予知システム全体の精度を算出することが出来る。また、システム全体から要予知日報判定部の影響を無くし、障害予知部単独の精度を測ることができる。

9 評価結果

9.1 要予知日報判定部の評価結果

評価実験 1 の結果から、要予知日報判定部の評価を行う。評価実験 1 で「障害が予知出来る」又は「障害が起こり得るが予知は出来ない」となった日報を正しく判定された日報として集計する。ここで集計された日報数を全日報数 (200 件) で割ることで要予知日報判定部の精度を算出する。算出した要予知日報判定精度を表 1 に示す。

表 1: 要予知日報判定部の精度

	要予知日報分類精度			
	評価者 1	評価者 2	評価者 3	平均
全ドメイン	0.685	0.585	0.595	0.622

表 1 に示したとおり、要予知日報判定部の精度は全体では 0.622 となった。

9.2 障害予知部の評価結果

評価実験 2 より、障害予知部のみの評価を行う。評価は日報を単位として用いる。日報単位の評価では、出力された予知情報のうち一つでも「起こり得る」と選択された場合、その日報に対して正しく予知できたとする。そして、正しく予知出来た日報数を入力日報数で割ることで精度を算出する。本節では、障害予知部のみの評価を目的としている。そのため、入力日報数とは評価実験 2 において「障害は起こり得ない」とされた日報を評価データから除いたものを示す。精度は、システムの出力、ランダムの出力の両方に対して算出する。障害予知部の評価結果を表 2 に示す。

表 2: 障害予知部の評価 (日報単位)

ドメイン	出力種別	要予知日報分類精度			
		評価者 1	評価者 2	評価者 3	平均
電話	システム	0.622	0.360	0.444	0.475
	ランダム	0.481	0.200	0.348	0.343
カメラ	システム	0.647	0.276	0.451	0.458
	ランダム	0.556	0.364	0.370	0.430
車	システム	0.639	0.286	0.500	0.475
	ランダム	0.594	0.200	0.478	0.424
ゲーム	システム	0.630	0.182	0.278	0.363
	ランダム	0.688	0.143	0.188	0.340
全体	システム	0.634	0.279	0.430	0.448
	ランダム	0.585	0.242	0.360	0.396

表 2 に示したとおり、評価者によって精度にばらつきがあるが、全てのドメインにおいて本手法による精度がランダムの出力による精度を越えることが出来ている。

10 考察

10.1 人間の予知能力について

本節では、評価実験 1 で得た「被験者が行った予知」を用いて人間の予知能力について考察する。評価実験 1 では、被験者に入力日報を見せ、自由記述で障害の予知を行わせた。この記述から、入力日報のうち何件に対して予知を行えたかを調査した。ここでは評価実験 1 において、「障害が予知できる」及び「障害が起こり得るが予知は出来ない」とされた日報を要予知日報とした。また、「障害が予知できる」とされた日報を予知可能日報とした。予知可能日報の要予知日報に対する割合を予知可能割合として算出した。調査結果を表 3 に示す。

表 3: 人間の予知可能割合

被験者	予知可能日報数	要予知日報数	予知可能割合
被験者 1	95	137	0.693
被験者 2	38	117	0.325
被験者 3	82	119	0.689
全体	215	373	0.576

表 3 より、人間は予知が必要な日報の六割弱に対して予知が可能であることが分かった。本システムの予知可能割合は表 2 より、0.448 である。この割合は、全体の予知可能割合を下回る結果ではあるが、評価者 2 の予知可能割合を越えることは出来ている。よって本システムの予知能力は、人間の予知能力を越えることが出来る可能性があることが分かった。

10.2 システムの予知と人間の予知の相違

本手法では、2 節で示したように、人間が気付きにくい障害を出力することを目的としている。よって、評価実験 1 で人間が行った予知と、システムの出力のうち起こり得ると評価された予知は異なるほど良い予知だと考える。

評価実験 1 で被験者が予知を記述し、かつ評価実験 2 でシステムの予知が起こり得ると評価された入力日報を考察対象として収集した。これらの日報を使って、システムの予知と人間の予知の相違を考察した。被験者が記述した予知は自由記述によるものである。機械的に一致しているかどうかを判断することはできない。よって、人間の予知とシステムの予知が一致しているかどうかの判断は人手で行った。考察した結果を表 4 に示す。「考察対象日報数」は評価実験 1 で被験者が予知を記述し、かつ評価実験 2 でシステムの予知が起こり得ると評価された入力日報数を示す。「一致日報数」はシステムの予知と人間の予知が一致した日報数を示す。

表 4 に示したとおり、一致日報数は考察対象日報数の 1 割程度となった。これは、システムの予知と人間の予知はほとんど一致しないということを示す。よって、本手法で行った予知は、人間が気付きにくい障害を多く出力できたと考える。

表 4: システムの予知と人間の予知の相違

	被験者 1	被験者 2	被験者 3
考察対象日報数	49	15	32
一致日報数	4	2	4

11 おわりに

入力された日報に対して起こり得る障害を予知し提示するシステムを構築した。手法は障害情報辞書の構築、日報と障害情報の対応付け、障害予知の三段階で構築した。本手法では日報と障害情報の対応付けを文書分類のタスクとして扱った。人手評価の結果、ベースラインとして用いたランダム出力の精度を越えることが出来た。また、システムの予知能力は人間の予知能力を越える可能性があることを示した。さらにシステムは人間の気付きにくい障害を多く予知することが出来た。

使用したツール及び言語資源

1) 形態素解析器 ChaSen, Ver.2.3.3, 奈良先端科学技術大学院大学 松本研究室, http://chasen.naist.jp/hiki/ChaSen/,
2) 構文解析器 CaboCha, Ver.0.53, 奈良先端科学技術大学院大学 松本研究室, http://chasen.org/~taku/software/cabocha/
3) 価格.com, 価格比較サイト http://kakaku.com/
4) EDR 電子化辞書, http://www2.nict.go.jp/r/r312/EDR/J_index.html

参考文献

1] Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama. Looking for trouble. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 185–192, 2008.

2] 乾孝司, 乾健太郎, 松本裕治. 接続詞「ため」に基づく文書集合からの因果関係知識の自動獲得. 情報処理学会論文誌, Vol. 45, No. 3, pp. 919–933, 2004.

3] Roxana Girju. Automatic detection of causal relations for question answering. *In Proc. of ACL 2003, Workshop on Multilingual Summarization and Question Answering - Machine Learning and Beyond*, pp. 76–83, 2003.

4] 価格.com. 価格比較サイト. http://kakaku.com/.

5] Suguru Aoki and Kazuhide Yamamoto. Opinion Extraction based on Syntactic Pieces. In *The 21st Pacific Asia Conference on Language, Information and Computation*, pp. 76–86, 2007.

6] 藤村滋, 豊田正史, 喜連川優. 文の構造を考慮した評判抽出手法. 電子情報通信学会第 16 回データ工学ワークショップ (DEWS2005), 2005. 6C-i8.

7] Alan Agresti and Brent A. Coull. Approximate is better than " exact " for interval estimation of binomial proportion. In *The American Statistician*, 第 52 巻.

8] Makoto Iwayama and Takenobu Tokunaga. A probabilistic model for text categorization:based on a single random variable with multiple values. In *In Proceedings of the 4th Conference on Applied Natural Language Processing*, pp. 162–167, 1994.

9] E. Milios, Y. Zhang, B. He, and L. Dong. Automatic term extraction and document similarity in special text corpora. In *Proceedings of the Sixth Conference of the Pacific Association for Computational Linguistics*, pp. 275–284, 2003.