構文片を用いた日報からの障害情報抽出

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

我々は、緊急な対応が必要である事実を表す表現を何らかの障害を報告している表現（障害情報）として定義し、これを自動的に抽出するシステムを提案する。

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。だが、管理する側の人間は全ての日報に目を通さなければならず、コストが非常に大きい。これは、大きな企業になればなるほど顕著である。また日報の数が膨大になれば緊急の対応が必要である日報の閲覧が遅れてしまう可能性がある。これは、大きく見れば企業の信用を失う事態に発展してしまう。もし、緊急の対応が必要である表現を抽出することが出来れば、閲覧コストを下げると共に、事態が大きくなる前に対応することが可能となる。また同表現を収集してリスト化し社員に公開することで、障害発生のリスクを低減することも可能となる。

2　関連研究

市村ら [1] は営業管理職向けの日報の閲覧コストの削減と意志決定支援を目的としたシステムを提案している。これは、知識辞書を用いて営業日報から成功事例・機会損失事例を抽出するものである。特定の企業の営業日報を元にしているためその企業に強く依存した辞書になっている。

斎藤ら [2] は障害管理情報から抽出ルールにしたがって障害の概要情報を抽出し可視化するシステムを提案している。抽出項目を「障害」「原因」「対策」とした場合の抽出精度はそれぞれ 0.878,0.701,0.703 となっている。プリンタに関する障害管理情報に限定して実験を行っているためドメインに大きく依存している。

両者とも学習データを元に辞書やルールを作成し入力からの情報を抽出している。これでは辞書作成のコストが大きいだけではなく、学習データに無い情報を抽出することが出来ない。また学習データに強く依存したシステムになっている。これらの問題を解決するため、我々は一般的な日報から自動的に辞書を作成し、辞書の拡張による未知の情報への対応を行っている。

3　障害情報の定義

本稿では障害情報を「ある日報の中で何らかの障害を報告している表現」と定義している。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。障害の内容を推察するには単語のみの表現では不十分である。例えば、「壊れる」という単語が障害情報として抽出されたとする。この単語のみでも何らかの障害が起こったことを予想することは出来る。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。障害の内容を推察するには単語よりも大きな単位が必要である。青木ら [3] は意見・評判情報の単位として構文片を提案している。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。構文片の抽出例を図 1 に示す。

構文片には以下の利点がある。

1. 抽出が容易である。
2. 統計情報が取りやすく扱いやすい。
3. 構文木に比べマッチングが容易である。
4. 意味のまとまりとして取り扱うことが出来る。

入力文：古いパソコンのバッテリーがいきなり爆発した。
出力構文片：　古い → パソコン
バッテリー → 爆発する
いきなり → 爆発する

図 1: 構文片抽出例

以上の利点は本稿で扱う障害情報にも当てはめることができると考える。よって本稿では障害情報を単語ではなく構文片で取り扱う。さらに障害情報は生じた障害の内容を推察することができる表現であるので、青木らの定義している構文片のうち「連用修飾」の構文片のみを障害情報として用いた。以下、構文片の修飾要素及び被修飾要素をそれぞれ前項、後項と呼ぶ。

4　提案手法

4.1　手法概要

図 2 に本手法の概要図を示す。

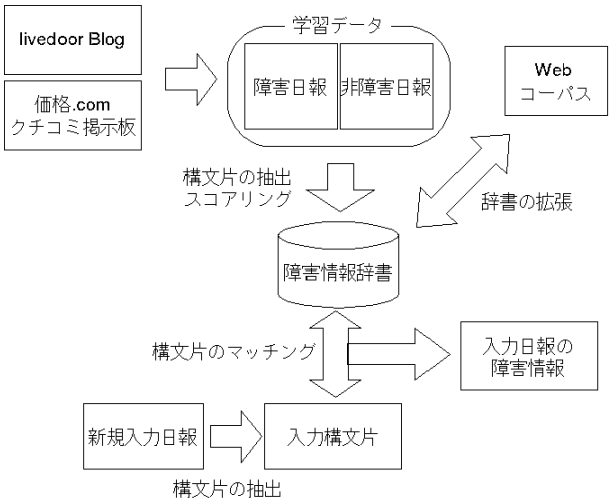

図 2: 手法概要図

本手法では、学習データの日報としてブログと掲示板を用いている。学習データから構文片を抽出しスコア付けを行い障害情報辞書を作成する。さらに学習データに存在しない障害情報にも対応するため障害情報辞書を Web コーパスを用いて拡張する。その辞書を用いて新規日報から障害情報を抽出する。以下に処理の詳細を示す。

4.2　学習データ

本稿ではブログや掲示板を一般的な日報の集合だと考え、学習データとして用いる。本稿で用いたブログは livedoor Blog[3] である。

このブログは 1 記事ごとにユーザーが独自のタグを付与することが可能である。また、1 記事ごとにタイトルを必ず入力しなければならない。そこでこれらの情報を基にタイトル又はタグに「トラブル」という単語が含まれている日報を障害情報を含んでいる日誌 (障害日報) とし、収集した。またタイトル及びタグ及び本文に「トラブル」という単語を含まない日報を非障害日報と

して収集した。

本稿で用いた掲示板は価格.com のクチコミ掲示板 [4] である。この掲示板はユーザーがスレッドを立てる時に話題のタグを選択するようになっている。本稿ではこの情報を用い、話題のタグが「悪い」に設定されているユーザーの発言を障害日報として収集した。また話題のタグが「悪い」「質問」以外に設定されているユーザーの発言を非障害日報として収集した。これらの日報を元に障害情報となる構文片を集めた辞書 (障害情報辞書) を作成する。

4.3 障害情報辞書の作成

収集した障害日報と非障害日報からそれぞれ構文片を収集した。得られた構文片に対して障害情報らしさのスコアを付与する。障害日報での構文片の出現回数と非障害日報での出現回数の差分を取ると一般的な意味で使用される構文片は絶対値が 0 に近付くと考えられる。また障害情報は障害日報での出現頻度が大きいと考えられ、正の値を持つはずである。よって、スコアの算出には藤村ら [4] の手法を参考にした。スコアの算出式を式 (1) に示す。

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

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

w_i はある構文片を表す。$P(w_i)$ はある構文片 w_i が出現した障害日報の数を示す。$N(w_i)$ はある構文片が出現した非障害日報の数を示す。P_{doc},N_{doc} は障害日報の総数、非障害日報の総数をそれぞれ表す。P_{doc},N_{doc} はその数に大きな差があるため式 (2) により母数を揃えた。式 (1) により算出したスコアが正の場合、その構文片は障害日報に出現しやすい構文片である。しかし、式 (1) では構文片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。例えば、出現頻度 100 で 0.9 のスコアを持つ構文片と出現頻度 1000 で 0.9 のスコアを持つ構文片は後者のほうがスコアの信頼性が高いはずである。この問題を解決するため我々は確率の信頼区間推定法 [5] を用いた。信頼区間を考慮した式を式 (3) にしめす。

$$score(w_i) = S'(w_i) - 2 * 1.96\sqrt{\frac{S'(w_i)(1 - S'(w_i))}{P'(w_i) + N(w_i) + 4}} \tag{3}$$

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

式 (3) の第二項が確率の信頼区間を示す。本稿では $score(w_i)$ が正の構文片のみ扱うため、負方向の信頼区間のみ考慮する。また式 (4) に示すように $score(w_i)$ は二つの確率の差分から求まる。よって信頼区間も双方を考慮しなければならないため式 (3) の信頼区間を 2 倍している。

以上より構文片にスコア付けし、正の値が与えられた構文片のみ障害情報辞書に登録した。

4.4 構文片を用いた障害情報辞書の拡張

新規入力日報からの障害情報の抽出は 4.3 節で作成した障害情報辞書とのマッチングにより行う。しかし辞書をそのまま用いると学習データに出現した障害情報しか抽出できない。そこで障害情報辞書を拡張し学習データにない障害情報にも対応した。以下に詳細を示す。

4.4.1 拡張の対象

拡張の対象の構文片はサ変名詞を含んだ構文片に限定した。さらに拡張部分はサ変名詞部分のみとした。これはサ変名詞以外の名詞部分が変化してしまうと、障害情報らしさも大きく変化してしまうと考えたからである。例えば、「メッセージが出ない」と「クレームが出ない」では前者には障害情報の可能性があるが、後者にはない。よって本稿ではサ変名詞以外の名詞は拡張対象としない。また動詞部分の拡張も本稿では行っていない。これは動詞には多くの活用形があり処理が複雑になることが予想されたからである。以後、拡張とはサ変名詞のみを対象としたものとして記す。

4.4.2 拡張方法

辞書を拡張する場合、拡張対象の構文片が持つ障害情報らしさを変化させてはならない。よって拡張対象のサ変名詞と同じ場面で用いやすいサ変名詞を得る必要がある。処理の概略を図 3 に示す。

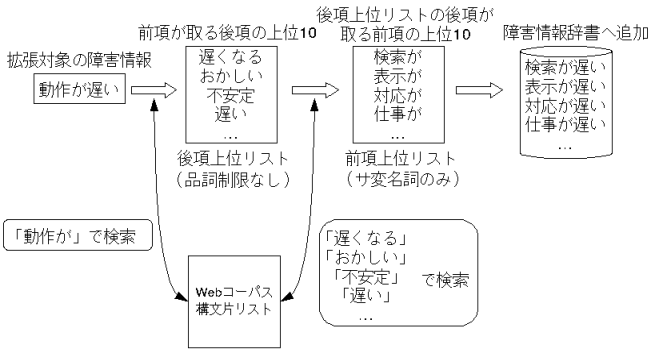

図 3: 拡張の概略図 (前項拡張)

まず我々は学習データとは別に収集した Web コーパスを用意した。このコーパスから構文片を抽出し構文片リストを作成した。このリストは Web コーパスに出現した構文片とその頻度を記録している。辞書の拡張は前項にサ変名詞を持つ構文片の拡張 (前項拡張)、後項にサ変名詞を持つ構文片の拡張 (後項拡張) の二つを行う。ここでは前項拡張の詳細を示す。

1. 拡張対象の構文片の前項で構文片リストを検索する。その前項が取る後項 (品詞制限無し) を収集し頻度の集計を行う。頻度の上位 10 件を後項上位リストとして得る。
2. 後項上位リストの後項 10 件を用いて再度外部構文片リストを検索する。10 件全てで検索しそれらの後項が取る前項 (サ変名詞のみ) を収集し頻度の集計を行う。頻度の上位 10 件を前項上位リストとして得る。
3. 前項上位リストに含まれる前項は拡張対象の構文片の前項と同じ場面で使われやすいものと考えることができる。よって前項上位リストの前項に拡張対象の後項を連結し障害情報辞書へ追加する。拡張して得られた構文片には拡張対象の構文片と同じスコアを付与する。

後項拡張については上記手順の「前項」と「後項」を互いに言い替えた処理を行い、障害情報辞書に登録した。

5 評価実験

4 節で構築した障害情報辞書を用いて新規の日報からの障害情報の抽出を行う。評価は二つの方法を用いる。

1. 二値分類器としての評価
   障害情報が抽出できた日報を障害日報、それ以外を非障害日報として障害・非障害の分類器を作成する。新規の日報に対して二値分類を行い分類精度を評価する。
2. 抽出された障害情報の評価
   新規の日報から抽出された障害情報を障害情報としての正しさで評価する。

5.1 評価データの作成

学習データとは別に用意した日報を人手で障害日報、非障害日報に分類した。分類は「障害を表す表現を含んでいるかどうか」

という基準で三人の被験者で行った。分類対象は日報のタイトルに「トラブル」という単語が入っている日報 400 件とタイトル及び本文に「トラブル」という単語が入っていない日報 200 件で行った。その結果三人一致で障害日報、非障害日報となったものがそれぞれ 133 件、253 件であった。ここで得た障害日報と同数の非障害日報を評価データとして用い評価実験を行った。

5.2 二値分類器としての評価

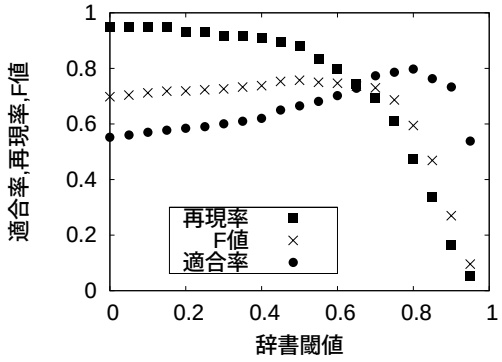

図 4: 二値分類器としての評価結果 (拡張なし)

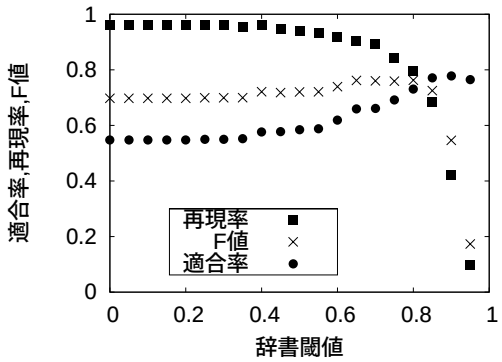

図 5: 二値分類器としての評価結果 (拡張あり)

図 4 は、4.3 節で構築した辞書を用いた結果である。二値分類器としての評価を適合率、再現率、F 値で行った。辞書内の構文片のスコアに閾値を設け閾値以上のスコアを持つ構文片のみ障害情報の抽出に使用している。F 値は閾値 0.50 以上の場合で最高値となりその値は 0.757 であった。またその時の適合率、再現率はそれぞれ 0.665,0.880 であった。図 5 は 4.4 節で拡張した辞書を用いた結果である。図 4 と同様にスコアに閾値を設けて障害情報の抽出を行っている。F 値は閾値 0.780 の場合で最高値となりその値は 0.772 であった。またその時の適合率、再現率はそれぞれ 0.724,0.827 であった。

5.3 抽出された障害情報の評価

本節では評価データから実際に抽出された障害情報を評価する。評価は 4.4 節で拡張した辞書を用いた結果に対して行う。辞書の閾値は 0.780 を用いた。図 5 で F 値がもっとも高くなった値である。この閾値で抽出された障害情報は 407 個であった。これらの障害情報を以下の基準で人手で評価した。

評価基準

(1) 何らかの障害を表している。

(2) 直接的に障害を表してはいないが何らかの障害を連想することが出来る。

(3) 障害を表しておらず、連想することも出来ない。

(1) の何らかの障害を表している表現は 116 個であった。(2) の障害を連想することが出来る表現は 47 個であった。(3) の障害を表しておらず、連想することも出来ない表現は 244 個であった。この結果から分かるように本手法は障害情報を得ることは出来るが適合率は 0.30 弱となった。(2) の表現まで正解だと考えると適合率は 0.40 程度であった。抽出された障害情報の再現率を正確に知ることは不可能である。しかし 5.2 節で示した二値分類器としての再現率は F 値最大の点で 0.827 であった。よって正確な値では無いが今回の障害情報の抽出は再現率 0.827 程度で行われていたと考える。この値は人手で作成した辞書やルールでは実現するのが難しい高い値である。

6 考察

6.1 精度について

本システムの評価は、5 節で示した通りの値となった。この値だけでは一概に良い結果だとは言えない。しかし、学習データの質を考慮すると本稿の手法は有効であったと考えられる。本稿で用いた学習データは 4.2 節に示したように簡単な基準で作成したデータであり、ノイズも多く含んでいると考えられる。実際 5.1 節で示したように、4.2 節で定めた基準で選んだ日報は被験者の三人一致のみが障害日報だとすると 400 件中 133 件という少ない数になった。被験者が一人でも障害日報だとした日報は 400 件中 227 件であった。逆に一人も障害日報だとしなかった日報は 400 件中 173 件であり、これは全て学習データ中のノイズだと考えられる。この割合は約 40 %でありこれだけのノイズを含んだデータで学習したシステムが 5 節で示した精度であったと考えれば本手法は今回の問題に対して有効であったと考えることが出来る。また、もしノイズをまったく含んでいない学習データが用意できたとすれば同じ手法でさらに高い精度を得ることも可能だと考える。

6.2 二値分類時の誤抽出数について

誤抽出とは、5.2 節で非障害日報を障害日報だとして抽出してしまったものを示す。図 6 に辞書の閾値を変化させたときの誤抽出数の推移を示す。

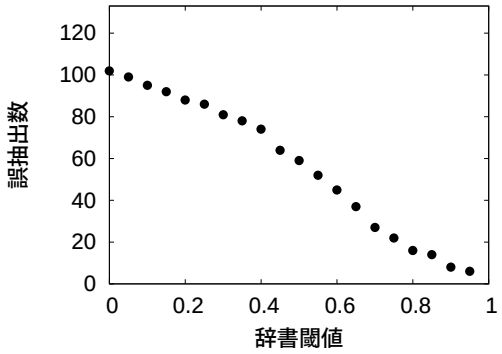

図 6: 二値分類時の誤抽出数の推移

図 6 より辞書の閾値を低下させると誤抽出数は線形に増加していることが分かる。これは辞書のどの閾値帯を選んでも同程度の割合で障害情報ではない構文片を含んでいることを意味する。理想的に近い辞書を構築出来たとしたら高いスコア帯には障害情報ではない構文片が少なく、低いスコア帯には同構文片が多くなるはずである。つまり誤抽出数のグラフは閾値の高い部分では傾きが小さく、閾値が低くなるにしたがって傾きが大きくなると考えられる。5.2 節と図 6 より、本稿で設定したスコア算出式 3 はスコア上位に障害情報らしい構文片を集めることが出来るが、どのスコア帯にも同程度に障害情報ではない構文片を含んでしまうということが言える。

6.3 抽出された障害情報について

評価データから抽出された障害情報の一部を表 1 に示す。表中の基準 (1)〜(3) は 5.3 節で定めた基準である。

表 1: 抽出された障害情報の例

基準	障害情報
(1)	画面-が ⇒ 表示-さ-れ-ない 遅延-が ⇒ 発生-する 音-が ⇒ 途切れる
(2)	サポート-に ⇒ 電話-する 販売-店-に ⇒ 返品-する 原因-を ⇒ 特定-する
(3)	コンセント-を ⇒ 抜く 電源-を ⇒ 入れる 一度-も ⇒ 繋がる

基準 (1) では否定の「ない」がつく表現が多くみられた。これより否定の「ない」を含む構文片は障害情報となりやすいということが言える。また、「音-が ⇒ 途切れる」は構文片という形になることによって障害の内容まで明確になっていることがわかる。基準 (2) は何らかの障害が発生したことにより出現した表現であると考える。実際に「サポート-に ⇒ 電話-する」は「画面-が ⇒ 表示-さ-れ-ない」と同じ日報に出現していた。これより同じ日報で出現した障害情報を関連付けて提示することにより基準 (2) も障害情報として見ることができ、障害の内容をより明確にする要因となると考える。基準 (3) は人が見ても障害だと言えない表現である。今回の実験では基準 (1)(2) の割合よりも基準 (3) の割合の方が多くなった。しかし、5.2 節での F 値は良好な値を示している。これより、人が見ても障害情報とは言えないものでも統計的に見れば障害日報に出現しやすく二値分類のタスクには有効な表現が存在することが言える。ただし今回の目的は障害情報の抽出であるので基準 (1)(2) と基準 (3) を区別することができるスコア付けを考案しなくてはならないと考える。

6.4 辞書の拡張について

表 2: 拡張で得られた障害情報の例

基準	障害情報
(1)	悪い ⇒ サービス → 悪い ⇒ イメージ 検索-が ⇒ でき-ない → 表示-が ⇒ でき-ない
(2)	サポートに ⇒ 連絡-する → サポート-に ⇒ 相談-する エラー-が ⇒ 出る → マーク-が ⇒ 出る
(3)	私-は ⇒ 報告-する → 私-は ⇒ 取引-する 連絡-を ⇒ くれる → 返事-を ⇒ くれる

表 2 に辞書の拡張によって得られた構文片の一部を示す。基準 (1) では「サービス → イメージ, 検索が → 表示が」基準 (2) では「連絡する → 相談する」「エラーが → マークが」が拡張によって得られた。これらは互いに似た語であることが分かる。また、拡張前の構文片の障害情報らしさを概ね保っていることが分かる。基準 (3) の例は「報告する → 取引する」「連絡を → 返事を」のように互いに似た語で拡張できている。しかし、拡張対象となった構文片が障害情報ではなかったため拡張した構文片も基準 (3) となってしまった。拡張した構文片によって得られた基準 (3) のほとんどが上記のような場合であった。これは構文片を用いた辞書の拡張自体は上手くいっていることを示す。よって、本稿で行った拡張方法は有効であったと考えることが出来る。本稿ではサ変名詞のみを対象に拡張を行った。サ変名詞のみに限定したため、実際に拡張された構文片は少数であった。拡張を動詞や一般名詞にまで適用出来る手法を考案することにより辞書の再現率を向上させることが出来ると考える。

7 今後の課題

今後の課題として三つの問題を挙げることができる。一つは学習データの収集についてである。今回示した収集法では 6.1 節で示したように約 40%のノイズを含んでいた。今回はタイトルに「トラブル」という単語を含んだ日報は全て障害日報とした。今後はタイトルだけでなく本文の単語も考慮しノイズの少ない学習データの収集法を考案する必要がある。二つ目は 5.3 節で示した基準 (2) の構文片の扱いである。何らかの障害を連想できるものの単体では障害情報とは言えない。他の構文片と組み合わせて提示することで障害情報として成り立つことが出来ると考える。最後に、障害情報の拡張についての課題が挙げられる。今回の拡張は 6.4 節より、有効であったことが分かった。しかしサ変名詞という制限からあまり多くの拡張は得られなかった。辞書の再現率を高めるため、今後はサ変名詞だけでなく動詞や一般名詞などの拡張も考慮する必要がある。

8 終わりに

我々は入力テキストから障害情報を自動的に抽出する手法を提案した。抽出には障害情報辞書を用いた。辞書は構文片という単位を用いて学習データより自動的に作成した。さらに学習データに無い未知の障害情報に対応するため、サ変名詞を対象に辞書を拡張した。構築した辞書を用いて新規の日報から障害情報を抽出した。評価は二通り行い二値分類器としての評価で F 値 0.772 を得た。また抽出された障害情報の評価で適合率で 0.40、再現率で 0.827 を得た。

謝辞

本研究は (株) ジムコとの共同研究として行った。

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) 係り受け解析器 Cabocha, Ver.0.5.3,
奈良先端科学技術大学院大学 松本研究室,
http://chasen.naist.jp/hiki/Cabocha/.
3) livedoor Blog, http://blog.livedoor.com/.
4) 価格.com クチコミ掲示板 ユーザーレビュー,
http://bbs.kakaku.com/bbs/.
5) SVM 学習ツール TinySVM,Ver.0.09,
奈良先端科学技術大学院大学 松本研究室,
http://chasen.org/˜taku/software/TinySVM/.

参考文献

1] 市村 由美, 中山 康子, 赤羽 俊男, 三好 みよ子, 関口 寿一, 藤原 庸祐: 日報分析システムの開発: 電子情報通信学会技術研究報告,NLC, 言語理解とコミュニケーション, Vol.100,No.401,pp.31-38

2] 斎藤 孝広, 渡部 勇: 障害情報からのマイニング: 情報処理学会研究報告,NL-142-20,2001.

3] S. Aoki and K. Yamamoto: Opinion Extraction based on Syntactic Pieces: Proc.of PACLIC21,pp.76-86,2007

4] 藤村 滋, 豊田 正史, 喜連川 優: Web からの評判および評価表現抽出に関する一考察: 夏のデータワークショップ, 3, 2,pp.57-60,2004.

5] Alan Agresti and Brent A. Coull.: Approximate is better than “exact” for interval estimation of binomial proportion.: The American Statistician,Vol.52,pp.119-126,1998.