2007－NL－181 (13)
2007／9／26

因果関係に着目した口コミ情報からの評判情報抽出

[1]高野 敦子 [2]池奥 渉太 [2]北村 泰彦
[1]兵庫大学 経済情報学部 経済情報学科
[2] 関西学院大学 理工学部

本稿は，ネット上に大量に記述された「口コミ情報」からの，有用な評判情報の自動抽出について述べる．「口コミ情報」の重要性が注目されているが，特に読者はその「理由」を重要視しているという報告がある．また，評価を示す表現は，同時に評価の具体的な内容を理由として示すことが多いことも知られている．我々は，ホテルに関する掲示板の情報を題材として，評判情報に現れる理由などの因果構造を分析し，構文解析結果を用いて表層的に因果構造を抽出できることを確認した．そのような評判情報の特性を用いることによって，既に認定された評価表現を使って，それと因果関係を成す評価表現を自動的に認識することができる．本稿では，この処理を繰り返すことによって，少数の評価種表現を基にして評価表現を学習しながら，有用な評判情報を自動的に抽出する手法を提案する．ホテルに関する掲示板の情報を用いて実験した結果，分野に特徴的な評判情報のみならず，個々の対象に個別的な評判情報が抽出できた．

Extracting Evaluative Expressions by Using Causal Relations

TAKANO Atsuko[1] IKEOKU Syota[2] KITAMURA Yasuhiko[2]
[1]Dept. of Economics & Information Science, Hyogo University
[2]Faculty of Science and Technology, Kwansei Gakuin University

We consider the problem of extracting evaluative expressions from very large amounts of Web documents. We have focused on importance of the reasons of evaluative expressions when readers refer to review written by others. Using hotel reviews as data, we proved that evaluative expressions often include the reasons that explain the evaluating contents concretely. Additionally, it is reported that causal relations as reasons can be identified automatically using the characteristics of in-text causal relations. By taking advantage of the characteristic, this paper proposes an automatic learning method to extracts useful evaluative expressions from a very small set of seed expressions and analyzing the causal structure of Web documents. In our experiment on discussion board messages about hotel, we could extract a good set of evaluative expressions relevant to the domain.

1．はじめに

近年,インターネット上に大量に記述される「口コミ情報」の影響力が注目されている．そして，その情報量が人手では処理できないほど大量になっている現在，有用な評判情報を自動的に抽出・解析するための技術への関心が高まっている．[1, 2, 3, 4, 5]

評判情報を構成する表現は多種多様である．さらに評価表現であるかどうか，あるいは好評なのか不評なのかは，評価する対象や観点に依存することが既に認識されており，分野毎に評価辞書を構築する試みなどが行われている．[2]

それに加えて本研究では，表現がどのような文脈で使われているかが評判情報の認識の上で重要であるとの考えの基に評判情報抽出の手法を検討する．例えば，

記事１「これで，部屋が綺麗なら最高なのに.」

という記事において，単純に辞書に登録された評価表現とのマッチングのみを用いて文脈を考慮せずに評判情報を抽出すると，“部屋が綺麗”という評判が抽出されるが，これはこの場合正しく認識したとは言えない.また,次の記事２において，

記事２「部屋は狭いが，スタッフの方の対応が親切だったので満足した.」

文脈を考慮しなければ，“部屋が狭い”，“スタッフの方が親切”，“満足する”という評判情報が同等に抽出されると考えられるが，“部屋が狭い”は少なくとも書き手が主張したい評判情報とは言えない.“スタッフの方が親切”，“満足”という評判情報がより重みのある情報と言える．

これに対して，映画のレビューにおける「理由」の重要性を分析した研究[7]では,『映画を見るかどうかの参考とするとき、「理由」を重視する』という調査結果が紹介されている．また，評価を示す表現は，具体的な理由として評価の内容を同時に示すことが多いことも知られている[6]．そこで，本研究では評価を示す表現とそれに対する具体的な評価内容の表現が評価の理由という因果関係を成す形で現れている評判情報を抽出することにより，より有用な評判情報の抽出を試みる．この因果関係を認識することにより，上述の記事１からは，“部屋が綺麗でない”,” 最高でない “という評判情報を抽出することが出来る．

また，次の記事３において，

記事３「三宮に近いので，気に入っている」

“三宮に近い”という記述は限定された対象に対して現れる評価であり，辞書に評価表現として登録されていない可能性が高い．しかし，“気に入る”という表現が評価表現であることがわかっていれば，“三宮に近い”がその理由を表す評価内容であることが推測される．この考え方を使って本研究では，対象となる口コミ情報に対して，評価表現辞書としては最初に，広く一般的に使われる少数の初期評価表現からなる辞書を用意するのみとし，記事内の因果関係を認識することにより新たに評価表現を学習しながら，自動的に有用な評判情報を抽出する仕組みを提案する．本稿では題材として楽天トラベルサイトの「お客様の声」を調査および実験に用いる．

2．関連研究

評判の抽出に関する先行研究としては，立石[1]らの研究が知られている．この研究では，あらかじめヒューリスティックな手法で構築した評価表現辞書を用いて評判情報を抽出している．辞書は分野毎に必要であると考えられるので，それらを人手で構築するためには大変な労力を必要とする．これに対し，小林らはこの評価表現辞書をテキストマイニングの手法を使って半自動的に強化する研究[1]を行っている．

評価表現獲得に手法として文脈一貫性を利用する研究としては，那須川らの研究[4]がある．評価表現の周囲には同じ極性を持つ評価表現が出現する可能性が高い，という性質を利用して，「満

足する」などの少数の評価表現からなる種表現を基に，ブートストラップ的に評価表現を学習し，自動で評価辞書を構築する．この研究が扱う「周辺に位置する表現」には本研究が対象とする「因果関係を持つ表現」も含まれる．ただし，那須川らの研究では，周辺に位置する表現に対して，出現頻度と極性の一致を用いてそれらを絞り込み，評価表現辞書の構築を目的とするため，分野毎に構築された辞書は各分野の特性は反映するが，個々の対象に対する評価の多様性までは考慮されない．例えば，特定の対象に限定して使われる評価表現や，「ホテル近辺が賑やか」のように，同じホテルの分野でも対象によっては好評であったり，不評であったりするような表現は抽出されにくい．これは評価辞書の構築を目的とした場合には自然な結果と言える．それに対して，本研究は，評価表現辞書として用いるのは小規模な初期辞書のみで，その都度評価表現を学習しながら評判情報を抽出する．ただし，1つの対象に対する文書の量が大量にあることは仮定する．周辺という位置関係だけでなく，構文解析結果を利用した因果関係の認識を図ることにより，文脈構造にまで考慮し，個別的な対象毎の評判情報の抽出を試みる．

3．因果構造を利用した評判情報抽出

3．1　因果構造に着目した口コミ情報の分析

評価を示す表現は，具体的な理由として評価の内容を同時に示すことが多いという特性が認識されていることを1章で述べたが，このことは本研究の重要な前提となるため，実際の口コミ情報を人手で分析する簡単な実験を行った．

楽天トラベルサイトのホテルに対する「お客様の声」[10]から典型的な評価表現である「良い」と「満足する」を含む832文を抽出した．それらを，因果構造の観点から以下の4タイプに分類した．この分類は本研究の調査目的を考慮し，恣意的な分類になっている．

タイプ名	因果構造
タイプ1	表層的に「AなのでB」　という因果構造を含み，内容的にもAがBの理由になっている．
タイプ2	表層的に「AならばB」　という因果構造を含む．
タイプ3	表層的には「AなのでB」　という因果構造を含むが，内容的にはAがBの理由になっていない．
タイプ4	因果構造を含まない．

表1．調査で分類した因果構造のタイプ

結果は，タイプ1が546文，タイプ2が28文，タイプ3が130文，タイプ4が128文であり，このことから，理由を伴う評判を抽出することが意味をもつことが確認できた．

また，このような因果構造は，表層的な情報を用いて自動的に判別できることが報告されている[8, 9]．判別方法の詳細は5章で述べる．

3．2　因果構造を持つ評判情報のモデル化

上記の結果から，本稿では，前節での分類の中で，理由を抽出可能な最初の2つのタイプの因果構造を持つ評判情報を抽出の対象とする．

【タイプ1】表現Aなので表現B

【タイプ2】表現Aならば表現B

ただし，タイプ2については，「～表現Aなので～表現B」と読み替える．

ここで，表現Bが対象に対する評価を表し，表現Aは，その理由として具体的な評価内容を表すと考え，表現Bを「評価表現」，表現Aを「評価内容」と呼ぶことにする．そして評判情報を本研究では，評価内容と評価表現の対としてとらえる．

1章で示した記事3：「三宮に近いので気に入っている」であれば，評価表現が「気に入る」であり，評価内容が「三宮に近い」となる．ここで，評判情報内での構造上の位置関係の観点から，評価表現と評価内容に分類したが，各表現が2つの分類のどちらに分類されるかは固定的ではなく，

したがって，この分類は排他的ではない．「建物が気に入っているので，満足している」という文においては，「気に入る」は評価内容として扱う．

3．3　因果構造を利用した評判情報抽出の基本的アイデア

本研究では，まず口コミ情報に対して，形態素解析・構文解析を行う．ここでは，形態素解析としてJUMAN,構文解析としてKNPを採用する．

種となる評価表現として「満足する」を取り上げてみる．

KNPで解析した結果が次のようになった場合，

料金が－↓
安かったので――|
かなり――↓
満足しました。

評価表現を「満足する」，評価内容を「料金が安い」とするタイプ1の因果構造を持つと認識できる．認識方法の詳細については5章で述べる．この認識の結果，「料金が安い」が評価内容候補として抽出される．

ただし，因果関係は表層情報のみを用いて判別するため，次のような，実際には理由になっていない関係も抽出する．

今回は－↓
観光だったので――|
とても――↓
満足しました。

このようにノイズも含んで抽出した評価内容候補に対し，文書全体の中で各候補表現が因果関係を持つ回数と全体で出現する回数に対する割合を計算し，それらの値に対して設定した判定ルールと照合して，基準を満たす候補を評価内容と認定する．このとき，認定された評価内容と，それと因果関係を持つ評価表現との対を評判情報として抽出する．

このようにして認定した評価内容を用いて，今度は評価表現を抽出する．例えば，

料金が－↓
安いので――|
有り難い。

という記事に対して，評価内容と認定されている「料金が安い」がタイプ1の因果関係の評価内容として現れた場合，「有り難い」が評価表現候補となる．この場合もノイズを含んだ評価表現候補が抽出されるため，文書全体の中で各候補表現が因果関係を持つ回数と全体で出現する回数に対する割合を計算し，それらの値に対して設定した判定ルールと照合して，基準を満たす候補を評価表現と認定する．この際も評価内容認定時と同様に，評価表現との対を評判情報として評判情報が抽出される．

このようにして評価内容と評価表現を交互に抽出する作業を何回か繰り返すことによって，この対象に限定された評判も含んだ，有用と考えられる理由付の評判情報を抽出することができるというのが本研究の基本的アイデアである．

ここで実際には，評価内容と評価表現に関して，好評・不評の極性の認定する必要があるが，それについてはさらに議論が必要なため，次稿に譲る．

3．4　評価内容と評価表現の定式化

前節までの説明では評価内容及び評価表現を，「料金が安い」というような自然言語の形で記述したが，実際に処理を行う上では，KNPを用いて構文解析した結果の部分木として扱う．頻度を計算する際に表現間で助詞等の表現の違いを吸収して同定するために，各ノードは付属語を除いて自立語のみで構成する．したがって，「料金が安い」は以下のような木構造で表現する．

料金　→　安い

評価表現の記述方法としては，(対象，属性（観点），評価値）の三つ組を採用する研究[2]が多いが，属性（観点）は必ずしも明示的に抽出できない場合も多く，本研究ではより柔軟な記述方法を採用する．

4．因果関係の認定

大規模なテキスト集合から自動的に因果関係を獲得する手法はいくつか提案されている[8, 9]．しかし未だ因果関係の出現特性について一定の知見は得られていない．そこで，ここでは，既に提案されている手法の考え方を基本として，KNPの解析結果から得られる情報を用いて因果関係を認定する緩い基準を策定した．緩い基準としたのは，口コミ情報は必ずしも正しい文法の則って書かれているとは限らないため，KNP の分析結果を厳密に分析することはあまり意味がないとの考えからである．緩い基準によって生じるノイズは，候補を絞り込む際に排除できると考えられる．

4．1　因果パターン

因果関係の表層パターンは構成単位や明示性の点から次のように分類できる．

【1】単文

①広い部屋で*くつろげた*．（理由が名詞句）

【2】複文

①部屋が広かったため，*くつろげた*．（明示的）

②部屋が広く，*くつろげた*．（非明示的）

【3】複数文

①部屋が広かった．おかげで*くつろげた*．（明示的）

②部屋が広かった．*くつろげた*．（非明示的）

この中で，本稿では，単文と複文から構成される因果関係を扱う．実験を重ねながら，複数文にも拡張していきたいと考えている．

4．2　解析結果からの因果関係抽出ルール

複文の場合，「ため，から，より，れば」などの接続助詞を用いた明示的な言語標識がある場合もあるが，実際の口コミ情報では，そのような言語標識の認められない連用修飾節を成す場合の方が圧倒的に多い．単文の場合は，名詞＋格助詞（で，が，・・・）という形で，格助詞が言語標識と成す場合もあるが，そのような言語標識を伴わない文も多い．

そこで，ＫＮＰの解析結果に含まれる，各文節の係り受け解析情報と品詞情報を用いて抽出ルールを以下のように策定した．ルールは，与えられた評価表現と因果関係を持つ評価内容候補を抽出するルールと，逆に与えられた評価内容と因果関係を持つ評価表現候補を抽出するルールの２種類からなる．

【評価内容候補抽出ルール】

与えられた評価表現の根またはその根が係る文節に係る文節で以下の(A)から(E)のいずれかの条件を満たす文節がある場合，その文節に含まれる自立語を根とし，その文節に係る文節の自立語から成るノードを葉とする部分木を評価内容候補として抽出する．

(A)<係:連用>　&&　(<用言:形> || <用言:動>)

(B)<連用要素>　&& <補文> &&　(<用言:形>|| <用言:動>)

(C)<連用要素>　&& <補文> &&　<体言>

(D)<係:連用> && <用言:判> &&　<体言>

(E)<連用要素> && <係:デ格> &&　<体言>

【評価表現候補抽出ルール】

与えられた評価内容の根が (A)から(E)のいずれかの条件を満たす文節に含まれるとき，その文節が係っている文節に含まれる自立語を評価表現候補として抽出する．

例えば，「良い」が評価表現と認定されている場合に「清潔感があるので良いと思う．」という文を解析した場合，下の図のように認識される．

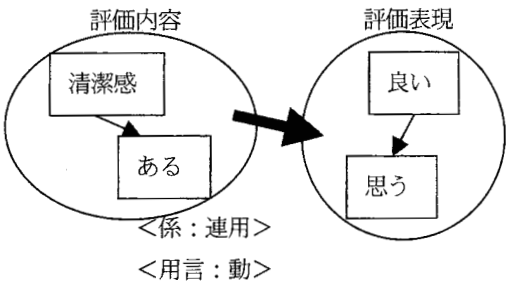

図１．因果関係抽出ルール

4．3　因果関係の判定基準

前節の抽出ルールを用いて抽出した評価内容候補や評価表現候補はノイズを含んでいるため，その出現分布を用いて絞り込みを行う．評価内容，評価表現と認定するための判定基準を次のように定める．

（1）候補認定回数（L）が十分多い（L＞K回以上）

（2）文書内全出現回数（T）に対して候補認定回数が十分多い（L／T＞M）

これらの基準における定数 K と M は実験を重ねながら調整する必要がある．

4．4　評価内容・評価表現同定方法

評価内容および評価表現の認定回数を数え上げるために，評価内容および評価表現の同定を行う必要がある．評価内容の場合には，直接評価表現に係る文節およびそれに係る文節を木構造として抽出するため．完全一致はしないが，同定と見なすことが望ましい木構造が認定される可能性がある．

例えば，「周辺に多数飲食店がある」，「飲食店が近くにある」など部分木が一致する木構造が認定された場合，別の表現として独立に認定回数をカウントすることは望ましくない．

そこで，部分木が一致する木構造が認定された場合は，共通する部分も合わせてカウントすることにする．

例えば，「飲食店→ある←近く」が N 回評価内容表現候補として認定されているところに，新しく「飲食店が周辺にある」が候補と認定された場合の表現同定方法は図２のようになる．

共通する部分木が N+1 回カウントされて候補に加えられる．

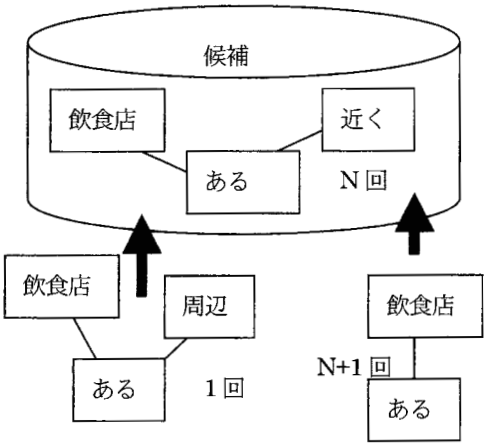

図２．表現の同定例

5．評価内容および評価表現抽出アルゴリズム

評判情報抽出処理の全体の流れは以下の図３のように，評価内容の抽出処理と評価表現の抽出処理を交互に繰り返すことになる．この過程で，評価内容，評価表現，およびそれらの対から構成される評判情報が漸進的に抽出される．

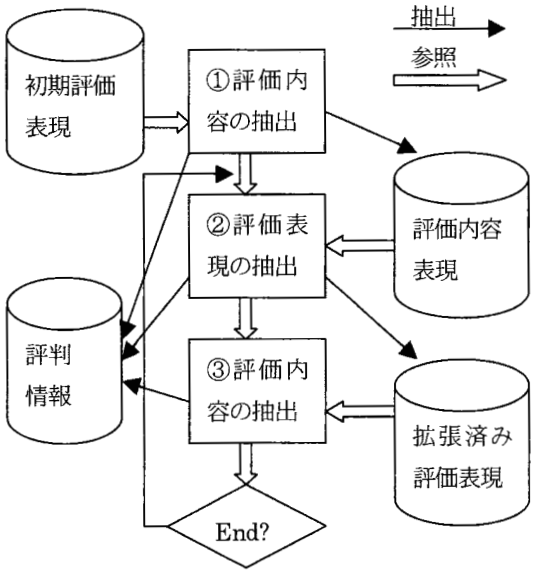

図３．全体の流れ

5．1　評価表現を用いた評価内容の認定

図３の①の部分の処理の流れは次の図４のよ

うになる．口コミ情報解析結果は，JUMAN による形態素解析，KNP による構文解析を行った結果である．「満足する」，「良い」などの広く一般的に評価に使われる少数の表現からなる（A）初期評価表現辞書を種として，4.2 節で述べた因果関係抽出ルールを参照して（B）評価内容候補を抽出する．これに対して，4.3 節で述べた因果関係判定基準を適用して（C）評価内容と評判情報を認定する．

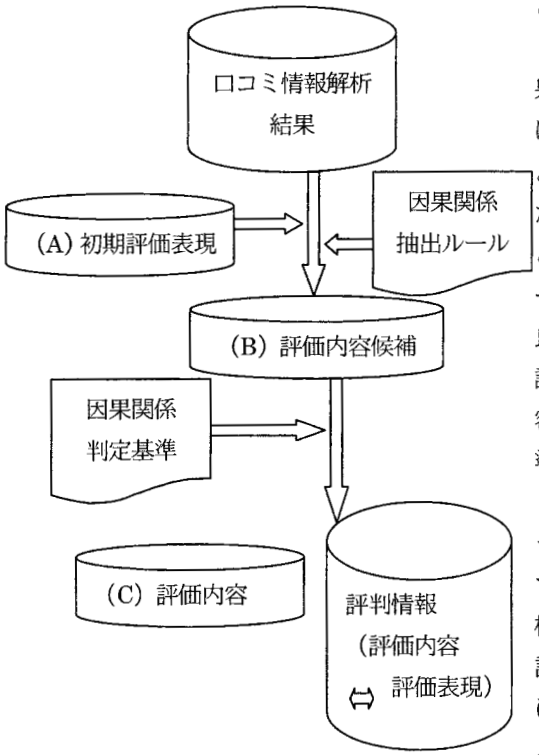

図４．初期評価表現を用いた評価内容抽出

図３の③の部分の処理の流れは次の図４において，(A) の初期評価表現が図３の②の処理で抽出した，評価表現を加えたものになる．

５．２　評価内容を用いた評価表現の認定

図３の②の部分の処理の流れは図４において，（A）が①の処理で抽出した評価内容に変わり，認定される（B）が評価表現候補となり，抽出される（C）評価表現と変わった流れになる．

６．実験結果

楽天トラベルサイトのデータ（15224 文）を対象として，種評価表現（「満足する」，「良い」，「いい」）から，評価内容と評価表現を抽出する実験を行っている．このデータは１つのホテルに対するデータである．まだシステムが完成していないため，本稿では定量的に有効性を検証することができないが，机上でのシミュレーションによる分析と合わせて，定性的な分析結果を述べる．

先ず，１回目の評価内容の抽出においては，因果判定ルールの基準を厳しく，ルール（１）における K の値を 10 に設定した場合，「駅→近い」および「品川駅→近い」が抽出される．に抽出できた表現を以下に示す．ここで既に「品川駅→近い」という対象固有の評価内容が抽出されている．ルールを緩めて K の値を５に設定すると，「立地→良い」，「部屋→広い」，「部屋→綺麗だ」といった評価内容が抽出され，評価内容と見なされない内容は抽出されなかった．K の値を２に設定して基準をかなりひくく設定すると，評価内容として，「東京タワー→見える」，「コンビニ→ある」といった多様な表現が抽出できた．K の値を１に設定すると，「トイレ→便座→温かい」などのさらに多様な表現が多数抽出できたが，書き込み情報の誤記や文法誤り，解析手法が対応できない表現などによって正しい構文解析が行われなかったことによるノイズも多数あった．しかしそれらは同じ表現が２回以上現れることはほとんどなかった．

上記の１回目の評価内容表現抽出において厳しいルールを採用して「駅→近い」と「品川駅→近い」の２表現を評価内容として評価表現を抽出した．評価表現は評価内容に比べて得られた表現の種類は多くなかった．閾値 K の値を 10 に設定して判定すると，「便利だ」と「利用する」が得られた．「利用する」は分野の特性を表す評価表現といえる．そのため，この２つの評価表現を加えて評

価内容を抽出すると，「安い」，「快適だ」，「リーズナブル」，「最高だ」，「プラン→ある」，などの一般的な評価内容及び分野固や対象固有の評価表現が多数抽出された．

7．まとめ

本稿では，評判情報に見られる因果構造に関する特性に着目することにより，少数の評価種表現を基にして漸進的に評価表現を学習しながら，有用な評判情報を自動的に抽出する手法を提案した．文脈情報を利用した研究では，文脈情報を使って抽出た候補に対し統計的な処理によって評価表現を絞り込むという方法がとられるが，先行研究との違いは，因果関係に注目することにより，最初からかなり絞り込んだ候補を抽出できることだと考えられる．そのために絞りこみの処理を緩い判定で行うことができるために，かなり対象に個別的な表現も抽出することが可能となる．また，本文中で述べたように，辞書を構築せず，毎回評価表現を学習するというやり方も，対象個別の評判情報の抽出を可能としている．

本稿では，実験システムが作成中であった為に、再現率，精度などの定性的評価ができなかった．この後定性的評価を行いながら，再帰的に繰り返す評価表現と評価内容の抽出処理の回数を増やすにつれて，抽出される表現がどのように変わっていくかを観察したいと考えている．

参考文献

[1] 立石健二, 石黒義英, 福島俊一. インターネットからの評判情報検索, 情報処理学会研究報告, NL-144-11, pp.75-82, 2001.

[2] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. テキストマイニングによる評価表現の収集. 研究報告「自然言語処理」No.154, 2003

[3] Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment Clasification using Machine Learning Techniques. Empirical Methods in Natural Language Processing(EMNLP2002) pp.76-86, 2002

[4] 那須川哲哉, 金山博 文脈一貫性を利用した極性付表現の語彙獲得, 情報処理学会研究報告, NL-162-16, pp.109-116, 2004.

[5]乾孝司, 乾健太郎, 松本裕治 ：出来事の望ましさ判定を目的とした語彙知識獲得，言語処理学会第 10 回年次大会発表論文集，2004.3.

[6] 那須川哲哉　テキストマイニングを使う技術／作る技術，東京電器大学出版社　5 章 pp.189-207, 2006.

[7]中山記男, 神門典子 レビューにおける「理由」の重要性の分析～被験者実験より～　情報処理学会研究報告, NL-171-14,pp.81-88, 2006.

[8] 乾孝司，奥村学　文書内に現れる因果関係の出現特性調査，情報処理学会研究報告, SLP-056,pp.81-86, 2005.

[9] 乾孝司, 乾健太郎, 松本祐治. 接続助詞「ため」を含む複文から因果関係知識を獲得する，情報処理学会研究報告, NL-150-25,pp.171-178, 2002

[10]　http://travel.rakuten.co.jp/