Vol.2014-NL-219 No.11
2014/12/17

# 接続標識「ノニ」をマーカーとした明示的でない因果関係の抽出

藤田 央[†1]　藤田 彬[†2]　田村直良[†3]

知的な情報処理の実現には，世界知識に限らず，知識源の表層に明示されない「常識的知識」を保持することが不可欠である．しかしながら管見の限り，文書から世界知識を獲得する試みは多くある一方で，常識的知識を獲得することに成功した例はない．常識的知識は書き手が読み手と共有されると認識した知識であり，文書上に示されることは少ない．本研究では，前提としてもつ推測や認識に対して意に反する事態が生じた際に書き手が認識の共有を確認する目的で常識的知識を表出する事例を抽出し，「因果関係」についての常識的知識を獲得する．不特定多数のユーザによる質問と回答を大規模に集積したコーパスを対象に，接続標識「ノニ」を含む文について「ノニ」の前後の因果関係を抽出する．「ノニ」の前後の因果関係を人手で判定し，獲得された知識の妥当性と手法の頑健性について，定量的，定性的に分析する．

# Extracting Implicit Causal Knowledge Using the Conjunctive Marker *noni*

HIROSHI FUJITA[†1]　AKIRA FUJITA[†2]
NAOYOSHI TAMURA[†3]

Not only the world knowledge but "the common-sense knowledge" which is implicit on knowledge sources is essential for the intellectual information processing. However, the acquiring of the common-sense knowledge has not succeeded yet. The common-sense knowledge is shared with readers and the writer. It is rarely shown on documents. In this study, we extract sentence samples which express causal common-sense knowledge from a corpus of a social Q&A service with the conjunctive marker “noni”. These samples were written to assure that the knowledge is shared between a writer and readers when the writer came across a situation against his/her assumption. We validate the appropriateness of the extracted causal common-sense knowledge.

## 1. はじめに

知的な情報処理の実現には，いわゆる常識的知識が必要である[1][3][8].

知識を計算機可読なデータにする研究には，1）教科書や専門書など専門的な文書に沿って人手で構築するアプローチと 2）言語データベースなどの電子データから計算機で抽出するアプローチの 2 種が挙げられる．前者は，広いドメインの知識を獲得することが困難な一方で，深い表現力を持つ知識ネットワークを構築できる．例として，世界史の教科書から人手で作る川添らのアプローチ[4][5]が挙げられる．後者は，様々な分野のコーパスから大規模に知識を獲得できる．後者のアプローチとして，新聞記事からある表現を手がかりに知識を獲得する例が挙げられる．乾らは接続標識「ため」に着目し[3], 坂地らは手がかり表現「を背景に」「に伴う」等に着目した[11].

書き手は自分の論理を展開するために必要な事柄を明示的に表出する．しかしながら，常識的知識は書き手が読み手と共有されうることを前提とするため，多くの場合表出されない．これまでの知識獲得の試みにおいては，文書上に明示的に表出される知識の抽出に成功した一方で，表層上の手がかりを捉え難い常識的知識については，管見の限り，適切な抽出プロセスを確立した例が存在しない．

本研究では，常識的知識として「一般的にはＡならば￢Ｂであるが，例外的にＢが成り立つ」(例「コンビニエンスストアで７５０円以上買い物したのに商品が当たるカードを貰えない」から一般的に「コンビニエンスストアで７５０円以上買い物したならば商品が当たるカードをもらえる」であるが例外的に「商品が当たるカードを貰えない」が成り立つ) 関係に関する知識(以下，因果関係知識)を抽出し，抽出した因果関係知識の妥当性を定量的に，抽出手法の頑健性を定性的に検討すること，抽出した因果関係知識を汎化することが可能かどうかを検討することを目的とする．

具体的には，投稿された質問と質問に対する回答の組を集めたコーパスを取り上げ，手がかり表現をマーカーとして常識的知識を抽出する．手がかり表現には，「意に反する事態」を表す際に記される接続標識「ノニ」を用いる．コーパスには，「意に反する事態」を集めたナレッジコミュニティサービス『Yahoo!知恵袋』を用いる．

†1 横浜国立大学環境情報学府
Graduate School of Environment and Information Sciences
Yokohama National University
†2 国立情報学研究所
National Institute of Informatics
†3 横浜国立大学環境情報学研究院
Graduate School of Environment and Information Sciences
Yokohama National University

## 2. 因果関係知識の抽出

### 2.1 抽出対象のコーパス

『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese)[6]に収録されている Yahoo!知恵袋サブコーパス[a]から因果関係知識を抽出する．

Yahoo!知恵袋は，ユーザの投稿質問に他のユーザが回答するナレッジコミュニティサービスである．ユーザが投稿する質問は「意に反する事態」が起こった時に書かれていることが多く， 因果関係知識を抽出する目的に適する．

Yahoo!知恵袋サブコーパス（以下，知恵袋コーパス）は，表 1 の 14 個の大カテゴリから構成される．知恵袋コーパスは，sample と呼ばれる質問本文と回答本文との組で構成される(付録 A.1)．sample の質問本文にある<sentence>タグと</sentence>タグで囲まれた部分(以下，質問文)を抽出の対象とした．sample の質問本文には，複数の質問文が含まれる．

表 1 Yahoo!知恵袋の大カテゴリ

| |
|---|
| エンターテインメントと趣味 |
| インターネット、PC と家電 |
| ビジネス、経済とお金 |
| 職業とキャリア |
| ニュース、政治、国際情勢 |
| スポーツ、アウトドア、車 |
| 暮らしと生活ガイド |
| 健康、美容とファッション |
| 子育てと学校 |
| マナー、冠婚葬祭 |
| 教養と学問、サイエンス |
| 地域、旅行、お出かけ |
| Yahoo!JAPAN |
| その他 |

### 2.2 抽出手順

本節では因果関係知識の抽出手順について説明する．

**Step1．MeCab によるフィルタリング**

知恵袋コーパスから抽出した質問文 217,238 文を質問文ごとに，MeCab[b]で形態素解析をし，接続助詞「ノニ」を含む質問文のみを抽出する．その際，解析結果に含まれる「のに　助詞,接続助詞,*,*,*,*,のに,ノニ,ノニ」を手がかりとする．

**Step2．二項関係部分が抽出できない質問文の削除**

残った質問文から「(従属節)＋のに＋(主節)」[c]の形(以下，二項関係部分)が抽出できない質問文を削除する．以下に削除する質問文の例を示す．

– 「文頭のノニ」が含まれる質問文や「文末のノニ」が含まれる質問文や「倒置文のノニ」が含まれる質問文を手作業で削除する．
  a. *なのに、夫はゲームにテレビ、映画など趣味三昧.* (文頭のノニ)
  b. *契約するつもりなかったのに.* (文末のノニ)
  c. *ばかだな　そら　なんておよげるわけないのに* (倒置文のノニ)
– 文意を解釈することができない質問文や，書かれていることに関する知識がないために作業者が文意を解釈できない質問文を手作業で削除する．

**Step3．因果関係知識である二項関係部分の抽出**

残った質問文から二項関係部分を手作業で抽出する．二項関係部分のうち，因果関係知識として扱わない二項関係部分を削除し，因果関係知識である二項関係部分を抽出する．以下に，因果関係知識として扱わない二項関係部分の具体例を示す．

  a. *前に住んでいた人はすごく静かだったのに今度の人はめちゃめちゃうるさいんです.* (対比)
  b. *観光するのにどのくらいの時間が必要ですか？* (目的)
  c. *爪先や踵が覆われている靴を履くのに抵抗を感じます.* (形式名詞「の」＋格助詞「に」)

**Step4．述語項構造の抽出**

日本語形態素解析システム JUMAN Ver.7.0[d]と日本語構文解析システム KNP Ver.4.11[e]を用いて，因果関係知識の従属節と主節それぞれから述語項構造を抽出する[12]．述語項構造を抽出する際に従属節の用言は終止形に変える．

### 2.3 因果関係知識データベース

抽出した因果関係知識を XML 形式のデータベースとして保持する．本データベースを加工して利用できる形式にするために，乾ら[3]，松吉ら[7]にならい，因果関係知識を自然言語のままで保持する．

因果関係知識データベースは，因果関係知識である二項関係部分を単位とする単位知識の集まりである．**単位知識**は，知識源に関する情報が書かれている**出典部**と，**因果関係知識部**との 2 つの部分からなる．因果関係知識部は，因果関係知識である二項関係部分をそのままの形で保持する**原文部**と，用言情報と述語項構造情報とが保持される**従属節情報部**，**主節情報部**との 3 つの部分からなる(図 1)．

- **用言情報**

<verbConcept>タグに，用言概念を保持する．用言概念は EDR 電子化辞書[9]の「動詞の概念」を用いた(表 2)．

a) 2004 年 10 月から 2005 年 10 月にかけて投稿された 3,120,839 の質問と，それに対するベストアンサーと呼ばれる回答の組の集合からサンプリングしたものが含まれる．
b) http://mecab.sourceforge.net/
c) 従属節，主節には名詞述語文も含まれる．
d) http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
e) http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

表 2 用言概念

| 概念名 | 説明 |
|---|---|
| 現象 | 現象(30f7e5)<br>自然現象(30f7e6)<br>静物に関する現象(30f7ea)<br>生理現象(30f7f4)<br>社会現象(30f7ff)<br>人間に関する現象(3aa947) |
| 行為 | 行為(30f83e)<br>自身行為(444d96)<br>対象行為(444dd8)<br>身体的活動(30f83f)<br>感情活動(30f863) |
| 移動 | 移動(30f801)<br>空間移動(30f802)<br>所有権の移動(30f826)<br>情報の移動(30f832) |
| 変化 | 変化(3f9856) |
| 状態 | 状態(3aa963)<br>性状・性向(3f9871)<br>物事に対する評価(30f7c8)<br>関係(30f9a4)<br>存在状態(3f98f6) |

※()は EDR 電子化辞書の概念体系辞書にある概念識別子

- **述語項構造情報**

<verb>タグに, KNP の解析結果の述語項構造 feature の代表表記を保持する．negation 属性に従属節，主節の用言の肯定/否定の情報を保持する．negation 属性の値は従属節，主節の用言が否定形であれば”有”と肯定形であれば”無”と記す[f]. 述語項構造 feature の格要素群の格に対応する<ga>, <wo>，<ni>，<de>などのタグに，述語項構造 feature の格要素群の表記を保持する．

## 3. 調査

### 3.1 因果関係知識の抽出過程と抽出数

2.2 に示す抽出過程のうち,「述語項構造の抽出」を除く，3 つの過程において抽出した質問文や二項関係部分の数を表 3 に示す．

**Step1** で自動的に質問文の 1.3%を抽出し，**Step2** で残りの 1.3%のうちの 55.5%を抽出し，**Step3** で残りの 88.2%を抽出した．因果関係知識の抽出過程における手作業の負担は少なく，有効な手法である．

表 3 抽出過程における質問文や二項関係部分の数

| | |
|---|---|
| 対象とした質問文 | 217,238 |
| MeCab によるフィルタリング | 2,915 |
| 二項関係部分が抽出できない質問文の削除 | 1,619 |
| 因果関係知識である二項関係部分の抽出 | 1,428 |

```
<?xml version ="1.0" encoding="UTF-8"?>
<commonSenseKnowledgeList >
    :
<commonSenseKnowledge number="20">
  <source>
    <corpus>Yahoo!知恵袋サブコーパス</corpus>
    <questionNumber>21</questionNumber>
    <sentenceNumber>1</sentenceNumber>
    <functionalExpression>のに</functionalExpression>
  </source>
  <causalKnowledge>
    <sentence>
      旦那が仕事で家にいるのに嫁は実家に帰省している。
    </sentence>
    <dependentClause>
      <verbConcept>状態</verbConcept>
      <verb negation="無" >居る/いる</verb>
      <ga>旦那</ga>
      <de>仕事</de>
      <ni>家</ni>
    </dependentClause >
    <independentClause>
      <verbConcept>移動</verbConcept>
      <verb negation="無" >帰省/きせい</verb>
      <ga>嫁</ga>
      <ni>実家</ni>
    </independentClause>
</causalKnowledge>
</commonSenseKnowledge>
    :
</commonSenseKnowledgeList>
```

図 1　因果関係知識データベースの XML 構造

### 3.2 因果関係知識か否かの判定での判定者間の一致率

判定者 2 名が二項関係部分 360 個[g]に対して「二項関係部分が因果関係知識であるか」を判定した．判定基準は，従属節(A)と主節(B)の間に「一般的にはＡならば￢Ｂであるが，例外的にＢが成り立つ」関係があると判定した場合は「因果関係知識である」，従属節(A)と主節(B)の間に「一般的にはＡならば￢Ｂであるが，例外的にＢが成り立つ」関係がないと判定した場合は「因果関係知識ではない」とした．また，両方にとれる場合は「因果関係知識である」とした．判定者間の「因果関係知識である」「因果関係知識ではない」の判定の kappa 値は 0.81 であった．判定者間の判定は十分に一致する．抽出した因果関係知識の妥当性が

f) 因果関係知識データベースを利用する際は，<functionalExpression>タグの情報をもとに，従属節と主節の用言が肯定形か否定形かを，ユーザが加工することを前提としている．

g) 調査に用いたに二項関係部分は，「因果関係知識である二項関係部分からランダムに選んだ 180 個」と「因果関係知識として扱わない二項関係部分からランダムに選んだ 180 個」との計 360 個である．

表 4　因果関係知識の分類

| | | 主節の用言概念 | | | | | |
|---|---|---|---|---|---|---|---|
| | | 現象 | 行為 | 移動 | 変化 | 状態 | 計 |
| 従属節の用言概念 | 現象 | 2 | 1 | 1 | 0 | 3 | 7 |
| | 行為 | 2 | 13 | 7 | 5 | 5 | 32 |
| | 移動 | 2 | 6 | 0 | 4 | 5 | 17 |
| | 変化 | 0 | 3 | 1 | 4 | 2 | 10 |
| | 状態 | 2 | 26 | 14 | 14 | 28 | 84 |
| | 計 | 8 | 49 | 23 | 27 | 43 | 150 |

確認された．

### 3.3 用言概念に着目した因果関係知識の分類

抽出した因果関係知識を汎化する際の手がかりとして，用言概念を用いて因果関係知識を分類する．分類は，「従属節の用言概念」，「主節の用言概念」の形式で表す．抽出した因果関係知識 1,428 個から無作為に 150 個を選び，因果関係知識の従属節および主節の用言を 5 つの概念: 現象，行為，移動，変化，状態に分け，因果関係知識を 25 分類した(表 4).

因果関係知識の分類の特徴として，以下が挙げられる．

- 25 分類のうち「状態，状態」，「状態，行為」，「行為，行為」「状態，移動」「状態，変化」の上位 5 分類で全体の約 60%を占めている．
  a. *同じ物件なのに家賃が違ったりする．*(状態,状態)
  b. *野球教室なのに野球を教えていない．*(状態,行為)
  c. *著作権者は著作権について騒ぐのに、他人が作ったホームページの内容を無断拝借する．*(行為,行為)
  d. *税金も支払済みなのに税務署の人が来る．*(状態,移動)
  e. *ドームでの試合なのに中止になった．*(状態,変化)
- 従属節の用言概念は，状態が全体の約 60%を占めている．主節の用言概念は，行為と状態がともに全体の約 30%を占めている．

## 4. 考察

### ● 因果関係知識の有効性

因果関係知識には，指示詞を含む事例が含まれる(1,428 個中，47 個)．指示詞を含む事例は，照応が解決されない限り，因果関係知識として不十分である．

### ● 判定者間で一致しなかった二項関係部分

判定者間で「因果関係知識である」「因果関係知識ではない」の判定が一致しなかった原因として以下が挙げられる．

- 解釈不能，または判断に必要な情報の不足
  例: *他のメーカーの蜂蜜ではならないのにこのメーカーの蜂蜜だけはなります.*
  →「ならない」が対象とする格要素が不明．
- 書き手の「思い込み」[h]
  例: *以前住友銀行でできたのに新銀行になってからはできなくなりました.*
  →「以前はできたのだから新しくなってもできる」と考えれば，「因果関係知識である」と判定され，「以前と今は無関係であり，書き手の思い込みである」と考えれば「因果関係知識ではない」と判定される．
- 読み手の知識不足
  例: *採用試験に落ちたのに履歴書を返してもらえない.*
  →読み手（判定者）が「採用試験における履歴書の取り扱い」について知識を持っていると「通例返却されない」ことを前提として判断できるが，知識を持たない場合は判定ができない．

これらの原因を解消するには「因果関係知識である」「因果関係知識ではない」の判定に関する基準を改善することが挙げられる．

### ● 因果関係知識の分類と汎化

抽出した因果関係知識を汎化するために，従属節と主節の用言概念によって分類した．しかしながら，因果関係知識が全 25 分類中上位 5 つの分類のみで全体の約 60%を占めている．分類の詳細化も含め，用言概念以外の分類による汎化を検討する必要がある．

### ● 因果関係知識と推論的因果関係

抽出した因果関係知識の中には推論的因果関係がある．下記の三文はある因果関係知識が別の推論的因果関係を使うことで説明できる例である．

(1) *他のものは臭くならないのに,1 枚だけ菌が根付いた.*
(2) *他のものは臭くならないのに，1 枚だけ臭くなる.*
(3) *1 枚だけ臭くなるのだから，1 枚だけ菌が根付いた.*

(1)は，(2)と(3)という知識が適用された因果関係知識である[2][10]．(1)(2)(3)を抽出した場合，(1)と(2)(3)が知識として重複する可能性があることに留意する必要がある．

### ● 因果関係知識と集団・場面との関係

Yahoo!知恵袋では，投稿された質問をカテゴリ別に整理している．因果関係知識を抽出するにあたって，どの集団での共有の知識なのか，どの場面での共有の知識なのかを留意する必要がある．

## 5. おわりに

接続標識「ノニ」をマーカーとして，知恵袋コーパスから因果関係知識を獲得する手法を検討した．

知恵袋コーパスから抽出した質問文 217,238 文に対し，MeCab によるフィルタリングで自動的に質問文の 1.3%を抽出し，二項関係部分が抽出できない質問文の削除で残りの 1.3%のうちの 55.5%を抽出し，因果関係である二項関係部分の抽出で残りの 88.2%を抽出した．最終的に 1,428 個の因果関係知識を抽出できた．抽出過程における手作業の

h) 書き手と読み手の間で共有されない暗黙的な知識．

負担は少なく，有効な手法である．

因果関係知識である二項関係部分と，因果関係知識として扱わない二項関係部分とを含む 360 個に対する，判定者間の「因果関係知識である」「因果関係知識ではない」の判定の kappa 値は 0.81 であった．抽出した因果関係知識の妥当性が確認された．

今後の課題として以下が挙げられる．

- 因果関係知識を汎化する手がかりとして，用言概念を用いて因果関係知識を分類することを試みた．しかしながら，表 4 に示したとおり，従属節，主節ともに，用言概念が「行為」，「状態」の頻度が高く，因果関係知識の分類の分布が均一ではない．用言概念「行為」，「状態」を細分化することで，因果関係知識の分類の分布が均一になり，因果関係知識の分類の「共通の性質，属性」が顕在化することが期待できる．
- Yahoo!知恵袋というナレッジコミュニティサービスに限定して，因果関係知識の抽出を行ったが，他のコーパスにも抽出手法が適用できるか検討するとともに，他のコーパスから抽出する因果関係知識が知恵袋コーパスから抽出した因果関係知識と異同はないか検討する必要がある．
- 「ノニ」以外の接続標識（たとえば「クセニ」「ニモカカワラズ」）を含む文に対し，本手法を適用し，明示的ではない因果関係知識を抽出することを検討する必要がある．

  a. *泣き虫のくせにけんかが強い.*

  b. *雨が降っているにもかかわらず傘をささない.*
- 抽出手順のうち手作業で行った過程に関して，削除ルールの知見が得られた．得られた知見を反映した因果関係知識の抽出支援システムにより，抽出にかかる時間が短縮され，抽出手法の頑健性と抽出する因果関係知識の妥当性とが向上することが期待できる．

## 参考文献

1) 新井紀子，松崎拓也：ロボットは東大に入れるか？―国立情報学研究所「人工頭脳」プロジェクト―，人工知能学会誌，Vol.27，No.5，pp.463-469(2012).
2) 有田節子：因果の言語学，月刊言語，Vol.25，No.5，pp.20-23(1996).
3) 乾孝司，乾健太郎，松本裕治：接続標識「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).
4) 川添愛，宮尾祐介，松崎拓也，横野光，新井紀子：「史実としてありえない」という判断を可能にする世界史オントロジー，人工知能学会第 27 回全国大会，2A4-5( 2013).
5) 川添愛，宮尾祐介，松崎拓也，横野光，新井紀子：出来事の成立・不成立の判断をサポートするイベントオントロジー，人工知能学会第 28 回全国大会，2I3-3(2014).
6) 国立国語研究所コーパス開発センター：『現代日本語書き言葉均衡コーパス』利用の手引 第 1.0 版，国立国語研究所コーパス開発センター(2011).
7) 松吉俊，村上浩司，乾健太郎，松本裕治：言語に基づく推論のための事象間関係知識データベースの構築，人工知能学会研究会資料，SIG-SWO-A802-08(2008).
8) 宮尾祐介，川添愛：「大学入試問題を解く」ことから見える言語，知識，世界理解に関する研究課題，人工知能学会誌，Vol.27，No.5，pp.470-478(2012).
9) 日本電子化辞書研究所：”EDR 電子化辞書仕様説明書”，日本電子化辞書研究所(1995).
10) 坂原茂：日常言語の推論，東京大学出版会(1986).
11) 坂地泰紀，竹内康介，関根聡，増山繁：構文パターンを用いた因果関係の抽出，言語処理学会第 14 回年次大会発表論文集，pp.1144-1147(2008).
12) 笹野遼平，河原大輔，黒橋禎夫，奥村学：構文・述語項構造解析システム KNP の解析の流れと特徴，言語処理学会第 19 回年次大会発表論文集，pp.110-113(2013).

# 付録

## 付録 A.1 「Yahoo!知恵袋」サブコーパスのサンプル例

| | |
|---|---|
| sample | 質問本文と回答本文を対にしたもの |
| OCQuestion | 質問本文を表す |
| OCAnswer | 回答本文を表す |
| sentence | 文に相当するまとまりを表す |

<?xml version="1.0" encoding="UTF-8"?>
<sample sampleID="OC12_04096" type="chiebukuro" version="1.0">
<OCQuestion>
<webLine>
<sentence>発展途上国の喫煙率が高い理由は何でしょうか？</sentence>
<sentence>お金がないのにタバコを買うお金はあるのでしょうか？</sentence>
</webLine>
<br type="logicalLine_original" />
</OCQuestion>
<OCAnswer>
<webLine>
<sentence>発展途上国においては喫煙を制限するような法的措置の欠如<br type="physicalLine_original" />喫煙による危険性の教育や情報が欠如しているのが現状です。</sentence>
</webLine>
<br type="logicalLine_original" />
<webLine>
<sentence>タバコ会社は先進国では売れないから規制の弱い発展途上国にタバコを売り始めました。</sentence>
</webLine>
<br type="logicalLine_original" />
<webLine>
<sentence>発展途上国の喫煙者は乏しい生活費の中からタバコを買い求めており<br type="physicalLine_original" />タバコ代が家計の圧迫することによる生活破壊が問題になっています。</sentence>
</webLine>
<br type="logicalLine_original" />
</OCAnswer>
</sample>