

特定の接続詞の意味特性を利用した 電子化文書からの因果知識の獲得法

A Method of Acquiring Causality Knowledge from Electronic Documents Using Semantics of the Specific Conjunctions

今給黎 勇佑†
Yuusuke Imakiire

石川 勉‡
Tsutomu Ishikawa

1. はじめに

コンピューターに人間のような柔軟な処理をさせるには常識知識の獲得とその活用が不可欠となる。常識知識のうち、事実知識に関しては、従来から百科事典や国語辞書からの獲得、インターネットを利用した不特定多数からの獲得等、広く研究されてきている[1-3]。これに対し因果知識の獲得については、いくつかの研究例[4,5]はあるがまだ広く研究されているとはいえない。常識的な因果知識は、冒頭の目的、特に柔軟な判断のための推論処理ではきわめて重要となる。ここでは、「にもかかわらず」等、特定の用法の接続詞を含む文からの因果知識の自動獲得について、先行研究[6]をもとに更なる高精度化を目指した検討結果について報告する。

2. 基本的な考え方

これまで、接続詞「ため」を用いた因果関係を持つ文からの知識の獲得について研究が行われてきた[4]。しかし、「ため」を用いた文には原因、目的、効果、前提といった複数の意味・用法が存在するため、それらを自動的に判別するのは非常に困難であった。

これに対し、接続詞「にもかかわらず」を用いた文は、以下の例のような二種類の関係に限定されるため、判別の自動化が容易であると考えられる。

例) ① 太郎は風邪にもかかわらず海に行った (因果)

② 太郎は勉強をしているにもかかわらず

次郎は外で遊んでいる (対比)

①の例では、「にもかかわらず」の前文の条件の下で、後文の内容は一般には起こらないことを表す。この用法で使われるときは、一般に“前文ならば後文ではない”ということの意味し、前文をP、後文をQで表せば、“P→Q”の形の知識に変換できる。従って、ここでは①の用法の文を知識獲得の対象とする。

ただし、後文Qが否定文の場合に上記のような知識変換を行うと“P→Q”となるが、この場合、PはQの必要条件でしかなく、因果知識としては適切ではないため知識変換の対象とはしない。

3. 変換法

先行研究[6]では、知識変換できるか否かの判定法として、①前文と後文の意味上の主語が同じである ②後文が否定文ではない、というふたつの条件を設定していた。

しかし、この条件に当てはまる文ならばすべて正しい

†拓殖大学大学院工学研究科電子情報工学専攻

‡拓殖大学工学部情報工学科

知識に変換できるとは限らず、間違った知識として出力されてしまうものがあつた。本報告では獲得精度を向上させるため、この条件をより細分化した。

3.1 前文Pが節の場合

前文Pが節の場合には、述語の品詞によって以下の2つの条件にあう場合は変換する。また、後文Qも基本的に節の形となるので、同様の変換法を用いる。

(1) 述語が動詞の場合

述語となっている動詞がEDR電子化辞書の日本語共起副辞書に記述されているかを調べる。記述があるならばその情報をもとに、入力文の動詞に必須格がすべてそろっているかを調べる。必須格がすべて揃っているならばそれらを用いて述語知識に変換する。この際、フィルモアの格文法を用いてラベル付けする。

例) 一郎は海に行く ⇒ 行く (agt:一郎, gol:海)

(2) 述語が名詞の場合

前文Pが“AはBである”という形の場合、Aは意味的には後文Qの主語である場合がある。入力文の係り受け関係を調べ、Aが後文に係っている場合、Aを後文の主語として扱い、後述の3.2節の変換法を用いる。

例) 僕たちは雨にもかかわらず遠足に行った

⇒雨にもかかわらず僕たちは遠足に行った

Aと後文に係り受け関係がない場合、Aはそのまま前文の主語として扱い変換する。シソーラス上でのAとBの上下関係を調べ、一定の上下関係がある場合は“inst”ラベル(述語の対象を表す)を用いて“B(inst:A)”と変換する。上下関係がない場合は、シソーラス上でAが属するカテゴリーを調べ補足語を補い、等値を表す“EQ”述語を用いて変換する。

例) 鯨は哺乳類だ ⇒ 哺乳類(inst:鯨)

太郎は介護者だ ⇒ EQ(atr:太郎, val:介護/者)

3.2 前文Pが単語(名詞句含む)の場合

前文Pが単語のみの場合や、“太郎の父”のような名詞句、“総合商社”のような複合語のみである場合、単純にはPを述語式には変換できない。そのため、さらに以下の3パターンに分類し、それぞれ補足語を補い述語式に変換する。

(1) Pが普通名詞の場合

Pが名詞のみの場合、まずその名詞がシソーラス上で抽象物に属するか具体物に属するかを調べる。抽象物であった場合、その単語の属するカテゴリーを調べ、補足語としてその上位概念を補い以下のようにEQ述語を用いて変換する。

例) 雨にもかかわらず花子は洗濯物を干した

⇒ EQ(atr:状況, val:雨) → 干す (agt:花子, obj:洗濯/物)

⇒ EQ(atr:状況, val:雨) → 干す (agt:x, obj:洗濯/物)

P が具体物である場合、以下の例のように前文の主語が省略されている可能性がある。

例) 多民族国家にもかかわらず

シンガポールは民族間の対立を制御してきたこの例の場合、前文は“シンガポールは多民族国家である”の省略形である。このように、後文の主語が前文 P の下位概念である固有名詞ならば、その単語を前文の主語として補い 3.1 節(2)と同様に変換する。

例) ⇒多/民族/国家(inst:シンガポール)→

→制御する(agt:シンガポール,obj:民族/間_の_対立)

⇒多/民族/国家(inst:x)→→制御する(agt:x,obj:民族/間_の_対立)

また、後文の主語が固有名詞でない場合でも、P の下位概念である普通名詞ならば同様に変換する。

例) カモノハシは哺乳類にもかかわらず卵を産む

⇒哺乳類(inst:カモノハシ)→→産む(agt:カモノハシ,obj:卵)

⇒哺乳類(inst:x)→→産む(agt:x,obj:卵)

(2) P が“普通名詞+サ変名詞”の名詞句・複合語の場合

P がサ変名詞で終わりそれ以外が普通名詞である複合語の場合、サ変名詞に「する」を付け加え動詞化する。普通名詞はサ変名詞の格として扱う。この普通名詞がサ変名詞に対して主格(“～が” “～は” という形の格)となる場合、知識として獲得する。

例) 景気低迷にもかかわらず市場が成長する

⇒景気が低迷するにもかかわらず市場が成長する

⇒低迷する(obj:景気)→→成長する(obj:市場)

(3) P が普通名詞で終わる名詞句・複合語の場合

複合語・名詞句をひとつの単語として扱い、(1)の形式で変換する。このとき、前に付く名詞は最後の名詞を修飾するものであるため、シソーラス上では最後の名詞と同じカテゴリとして扱う。

例) 集中豪雨にもかかわらず太郎は海に行った

⇒EQ(atr:状況,val:集中豪雨)→→行く(agt:太郎,gol:海)

⇒EQ(atr:状況,val:集中豪雨)→→行く(agt:x,gol:海)

4. 知識獲得量の増加

3章で述べた処理により、知識変換の正解率は上昇した。しかし、これは誤りのある知識を除外するために、正しく変換されていた知識も一部除外してしまい、結果として知識獲得量は減少した。これを補うため、「にもかかわらず」と同様の意味・用法を持つ接続詞に着目した。

具体的には、接続詞「なのに」を用いる。この接続詞は、断定の意味を持つ助動詞「だ」の連体形に、逆説の接続詞「のに」がついた形になっている。接続詞「のに」は広辞苑には“ある事柄から普通に予期されたことと反対の事柄が起こったことを示すのに用いる”とあるが、実際には“勉強をするのに遅すぎるということは無い”のように、これ以外の用法も存在する。これに対し接続詞「なのに」は、助動詞「だ」が付くことで前文が断定形に限定され、逆説の接続詞「のに」により後文は前文と相反する内容になる。これは「にもかかわらず」と非常に近い用法であり、同様の変換処理が使えらる。と考えられる。

接続詞「なのに」を含む文を毎日新聞 1994 年全データから 1164 文抽出し、そのすべてに対し「なのに」⇒「にもかかわらず」の置換を行ったところ、「なのに対し」「なのに比べ」といった特定の表現の文(2章で述べた“対比”の用法にあたる)を除き、すべてが日本語として

誤りの無い文になった。よって、「なのに」を含む文からも知識獲得を試みる。

5. 実験と評価

5.1 変換精度の検証

3章で述べた処理により、変換精度や取得率がどの程度変わるかを検証した。検証には毎日新聞 1994 年全データを用いた。毎日新聞 1994 年の記事全文には「にもかかわらず」を含む文が 780 個あった。そのすべてに対して、従来システムと本報告の新システムで変換処理を行った。従来システムでは 780 文から 445 文を知識として変換し、そのうち 244 文が正しい知識であった(正解率 55%)。新システムでは 132 文を知識として変換し、そのうち 105 文が正しい知識であった(正解率 80%)。

5.2 獲得量の検証

4章で述べた、接続詞「なのに」の利用により、知識の獲得量がどの程度増加するかを 5.1 で用いたデータにより検証した。この中には「なのに」を含む文が 1164 個あった。そのすべてに対し変換処理を行ったところ、99 文を新しい知識として獲得することができた。

以上の検証結果を表 1 にまとめて示す。

表 1: 各接続詞ごとの獲得量と正解率

	獲得量	正解数
従来システム (にもかかわらず)	445 (57%)	244 (55%)
新システム (にもかかわらず)	132 (17%)	105 (80%)
新システム (なのに)	128 (11%)	99 (78%)

6. まとめ

接続詞「にもかかわらず」を含む文から因果知識を獲得する方法について研究した。本手法は「にもかかわらず」の前文(P)と後文(Q)から、“P→Q”の形の知識を述語論理形式で獲得するものである。

従来の手法を見直すことにより、誤った出力の減少に成功した。また、それに伴い減少した獲得量を補うため、接続詞「なのに」を含む文に対し、「にもかかわらず」と同様の変換処理を適用する方法を提案した。

参考文献

- [1] D.B.Lenat,R.V.GUha,K.Pittman,D.prattand M.shepherd:Cyc “Toward Programs with Common Sense” Communications of the ACM,Vol.33.No.8(1999)
- [2] マインドピクセル,http://www.mindpixel.com/
- [3] 佐々木智彦,増子公徳,野中昌行,石川勉“国語辞書からの常識知識の獲得”,情報処理学会第 65 回全国大会講演文集(2), pp137-138(2003)
- [4] 乾孝司,乾健太郎,松本裕治:“接続標識”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,Vol45, No3, pp919-933(2004)
- [5] 佐藤浩史,笠原要,松沢和光,“表層的因果知識ベースによる事象推移予測方式”,情報処理学会第 56 回全国大会, Vol.2, 4Q-9, pp251-252(1998)
- [6] 富川昌則,石川勉:“接続詞「にもかかわらず」を含む文からの因果知識の獲得”,情報処理学会第 67 回全国大会講演文集(2), 1R-1, pp323-324(2004)