E-048

同一文中に出現する複数の節間における因果関係抽出の検討

A Study on Causal Relation Extraction between Two Clauses in a Sentence

山田 一郎† 宮崎勝† 三浦 菊佳† 住吉英樹† 八木伸行†
Ichiro Yamada Masaru Miyazaki Kikuka Miura Hideki Sumiyoshi Nobuyuki Yagi

1. はじめに

ディジタル放送では、データ放送や番組のクローズドキャプションなど大量の信頼できるテキストデータが放送波に多重されている。受信機が、このテキストデータから、有益な情報を抽出・蓄積できれば、視聴者からの様々な質問に答える賢いテレビが実現可能と考えられる。そこで我々は、クローズドキャプションを対象として、番組で表現された事柄の因果関係を自動抽出する研究を進めている。この因果関係は番組知識として映像を利用した質問応答などのアプリケーションに利用可能となる。本稿では、同一文中に出現する 2 つの節が属する意味カテゴリーを特定することにより、節間の関係を推定する手法を提案する。健康番組において顕著に出現する関係として“ 原因”、“ 症状”、“ 目的”、“ 対処法”の推定を試みた。NHK で放送された番組「きょうの健康」のクローズドキャプションを処理対象とした節間の関係推定処理と実験について報告する。

2. 関連研究

同一文中における因果関係の表現は、以下の 2 種類に分類できる。

(1) 名詞のペアに因果関係がある場合
例) 脳卒中(結果)の原因となる動脈硬化(原因)が促進される。
(2) 節のペアに因果関係がある場合
例) 急に運動を始める(原因)と血圧が急上昇します(結果)。

我々はこれまでに、(1)に示す名詞ペアを対象として因果関係の有無を判定する手法を提案している[1]。この手法では、名詞ペアの語彙情報と、名詞ペアの共通係り先までの単語情報を特徴として利用し、EM アルゴリズムによる判別実験により一定の分別能力があることが示された。 (2)に示す節のペアに対する因果関係を抽出する従来研究として、乾ら[2]は因果関係を“ 原因”、“ 効果”、“ 前提条件”、“ 手段”の 4 つに分け、「ため」という単語を手掛かり語として抽出した因果関係にある 2 つの節が、いずれに属するかを推定する手法を提案している。しかし一般的な文章中では、接続標識「ため」を利用して明示的に因果関係を表現する頻度は少ない。鳥澤[3]は、並列句の関係にある 2 つの動詞が共通の目的語を持つ時に因果関係が成立しやすいと仮定して、統計的に因果関係知識を抽出する手法を提案している。この手法は「ビールを飲む（原因）」→「ビールに酔う（結果）」といった常識的な因果関係抽出に有効であるが、(2)に示すような専門的な知識に関する因果関係では、利用される動詞に共通の目的語が少なくなり精度の低下が予想される。本稿では、同一文中の節を対象として、「ため」などの明示的な手掛かり語が無い場合における、医療事項に関する専門的な知識となる節間の関係を推定することを目的とする。

† NHK 放送技術研究所

3. 節間の関係推定処理

節間の関係を推定するために、まず、クローズドキャプションから処理対象となる節のペアを抽出する。次に、抽出した節がどのような意味カテゴリーに属すかを手作業で与えたルールにより分類する。適切な意味カテゴリー分類ができれば、節のペアが属する意味カテゴリーの組み合わせにより節間の関係が推定できると仮説を立て、フィッシャーの正確確率検定[4]による検定を行う。検定の結果、節のペアが属する意味カテゴリーの組み合わせが顕著に現れる関係が判明する。この節のペアが属する意味カテゴリーと関係を利用することにより、テストデータから節の関係を推定する。以下に各処理の詳細を記す。

3.1 節ペアの抽出

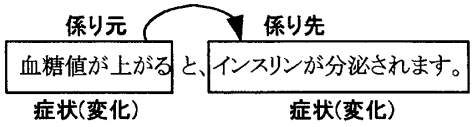
複文において、節に含まれる述語を表す文節が、別の節に含まれる述語を表す文節を修飾する場合、この 2 つの節は「原因－結果」などの関係を持つ可能性がある。そこで、係り受け関係にある節のペアを抽出する。この時、以下に示す 2 つの条件満たす節ペアに処理対象を制限した。

- 係り元の述語中の動詞が連用形である、または接続助詞「て」、「と」、「ば」を付属語として伴う
- 節ペアのいずれかに健康関連に特有な単語が含まれる

健康関連に特有な単語は、あらかじめ TFIDF 値を計算し、この値の上位を利用した。この処理により、節のペアが大量に抽出される。

3.2 節の分類

名詞ペアの関係を判定する処理では、名詞ペアの共通係り先までの単語列が重要な情報となるが、節ペアの場合、直接係り受け関係にあるため節の周辺情報は利用できない。そこで、節に含まれる情報を解析する必要がある。節が意味する事柄によって、その関係が特定できる場合がある。例えば以下の例では、係り元の節と係り先の節がともに「病状の変化」を表す。このようなケースでは係り元の節が係り先の節の原因を表すことが多い。



そこで、処理対象として抽出した節を以下に示す 8 種類の意味カテゴリーに分類する。

【節の意味カテゴリー】

- 症状（状態） 例）血圧が 高い
- 症状（変化） 例）細胞が 障害を 受ける
- 病気（状態） 例）糖尿病が 続く
- 病気（変化） 例）肺炎に なる
- 行為（医療） 例）インスリンを 注射する
- 行為（体の動作） 例）ひざを 伸ばす
- 行為（管理） 例）血圧を コントロールする
- その他 例）糖尿病を 含める

分類のために、節に含まれる動詞とその格構造の組み合わせを基とするルールを手作業により作成した。作成したルールの一部を以下に示す。

【ルール例】

[病気]の＋[*]が＋ 起きる → 病気（変化）
[内臓 or 分泌物]が＋不足する → 症状（状態）
[内臓 or 分泌物]を＋取る → 行為（医療）
[症状]を＋保つ → 行為（管理）

このルールにおいて、[病気]は、「病気」のカテゴリーに属する名詞を示す。このカテゴリーは、あらかじめ「病気」「症状」「行為」「人体属性」「内臓 or 分泌物」「体の部位」「医療品」などに属する名詞を人手により登録したものを利用する。[*]は任意の単語との一致を許す。

3.3 節間の関係の分類

番組のクローズドキャプションには、出現する節の間に様々な関係が存在する。提案手法では健康に関する番組について、以下の 4 つの関係とその他を分類対象とする。

- 原因（係り先の節の原因が係り元の節）
  例）インスリンの分泌を増やして、血糖を下げます
- 症状（係り元の節の症状が係り先の節）
  例）糖尿病になると腸管からコレステロールの吸収が増え、…
- 目的（係り元の節の目的が係り先の節）
  例）生活習慣を変えて内臓脂肪や肥満を取る
- 対処法（係り元の節の対処法が係り先の節）
  例）腎不全になり最終的には透析を行う…
- その他（上記 4 つの関係以外）
  例）眼科へ来て、初めて糖尿病が分かるケースが…

3.4 フィッシャーの正確確率検定による判定

節間の関係に特徴的な節ペアが属する意味カテゴリーの組み合わせを判定するために、フィッシャーの正確確率検定を用いる。フィッシャーの正確確率検定は、2 つの変数の間に統計学的に有意な差があるかを判定する検定手法で、近似せずに全ての可能な事象について列挙し、直接有意確率を計算する。節間の関係 x と節の意味カテゴリーの組合せ A との関係を考える場合、以下に示すような 2x2 分割表を作成する。

	関係x	関係x以外	計
意味カテゴリーA	a	b	$a+b$
意味カテゴリーA以外	c	d	$c+d$
計	$a+c$	$b+d$	$a+b+c+d$

この事例が出現する確率 p は以下の式で与えられる。

$$p = \frac{{}_{a+b}C_a \times {}_{c+d}C_c}{{}_{a+b+c+d}C_{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!}$$

節間の関係 x と節の意味カテゴリーの組み合わせ A に有意な差があるかを片側検定により判定する場合、以下の式により頻度 a 以上の確率値の和を求める。

$$\text{有意確率} = \sum_{\alpha \geq a} p(\text{意味カテゴリー}A\text{と関係}x\text{の共起頻度}=\alpha)$$

この有意確率が一定値以下の場合、節の意味カテゴリーの組み合わせ A は関係 x を持つ場合と判定できる。

4．実験

NHK で放送された「きょうの健康」の糖尿病に関連する 80 番組を処理対象とし、抽出した節の分類実験と、節間の関係推定実験を行った。処理対象のクローズドキャプションを解析して節ペアを抽出し、節の属する意味カテゴリーと、節間の関係の正解を人手により与えた。このうち半分を学習用データ、残りの半分をテスト用データとし、学習用データのみを参照して人手により 3.2 節に述べた節を分類するためのルールを作成した。このルールを用いて、テスト用データにある節を 8 つの意味カテゴリーに分類し、その他以外をまとめて評価した結果を以下に示す。

適合率	再現率
98.0% (343/350)	72.7% (343/472)

節の意味カテゴリー分類結果を利用して、学習用データから節間の関係に特徴的な節ペアが属する意味カテゴリーの組み合わせを抽出した。5%の有意水準による検証を行い、節の意味カテゴリー（係り元と係り先の組み合わせ）と節間の関係 11 組を抽出した。有意確率の値の小さい 5 項目を以下に示す。

係り元	係り先	節間の関係	有意確率
症状(変化)	症状(変化)	原因	2.1e-12
病気(変化)	症状(変化)	症状	4.1e-9
行為(医療)	行為(医療)	目的	3.1e-7
行為(体の動作)	症状(変化)	原因	4.8e-4
病気(状態)	行為(管理)	対処法	9.3e-4

抽出した節の意味カテゴリーと節間の関係を利用し、テスト用データから節間の関係を推定した。原因、症状、目的、対処法の 4 つの関係をまとめた評価結果を以下に示す。

適合率	再現率
81.0% (51/63)	31.3% (51/163)

節の意味カテゴリー分類結果の再現率が低いため、節間の関係推定実験の再現率は低いが、適合率は 80%を超え、一定の分別能力があると判断できる。

5. まとめ

本稿では、同一文中に出現する 2 つの節に対して推定した意味カテゴリーを利用することにより、節間の関係を推定する手法を提案する手法を提案した。実験により、健康に関する知識の一部を自動獲得できることを確認した。今後、節間の関係分類における再現率の向上を目指すとともに、健康に関するユーザからの質問に、映像で回答する機能を持つマルチメディア健康百科事典[5]へと応用していく予定である。

【参考文献】

[1] 山田ほか: クローズドキャプションを対象とした因果関係知識抽出の検討, FIT2005, E001, pp113-114(2005)
[2] 乾ほか: 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, 情処論 Vol.45, No.3, pp.919-933(2004)
[3] 鳥澤:「常識的」推論規則のコーパスからの自動抽出, 言語処理学会第 9 回年次大会, pp.318-321(2003)
[4] William L. Hays, "Statistics: Analyzing Qualitative Data," Rinehart and Winston, Inc., Chapter18, pp769-783(1988)
[5] 宮崎ほか: 番組字幕を利用したマルチメディア健康百科事典構築に関する検討, FIT2007, 4H-2, (2007)