放送番組のクローズドキャプションを対象とした健康に関する知識獲得へ向けて

山田 一郎　　宮崎勝　　三浦 菊佳　　住吉英樹　　柴田正啓　　八木伸行

ＮＨＫ放送技術研究所
E-mail: yamada.i-hy@nhk.or.jp

1. はじめに

デジタル放送では、データ放送や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、有益な情報を知識として抽出・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。本稿では、番組のクローズドキャプションから健康に関する事柄間の関係を自動抽出し、知識として蓄積する手法を提案する。まず、健康に関する事柄を表現する節を抽出する。次に、各節が属する意味カテゴリーを特定することにより、同一文中に出現する 2 つの節が“原因”、“症状”、“目的”、“対処法”の関係を持つか推定する。また、抽出した節には必要な情報が省略され、どのような事柄を表現しているか曖昧な場合がある。そこで、健康に関する事柄を表現する節の曖昧性を評価する手法も提案する。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実験について報告する。

2. 関連研究

文中に出現する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[1]。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[2]。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、健康や医療などの専門的な知識に関する因果関係では、使われる動詞に共通の目的語が少なく精度の低下が予想される。例えば、「風邪をひく（原因）」→「関節が痛くなる（結果）」という因果関係では、2 つの動詞が共通の目的語を持つことは少ない。本稿では、同一文中の節を対象として、明示的な手掛かり語が無い場合における、健康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。

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。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[3]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、節が別の節を修飾する場合、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の話題に特有な単語が含まれる

健康の話題に特有な単語は、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従来手法における処理[1]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して利用する。節間の関係は、2 つの節が持つ意味内容によって特定できる場合がある。例えば図 1 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。

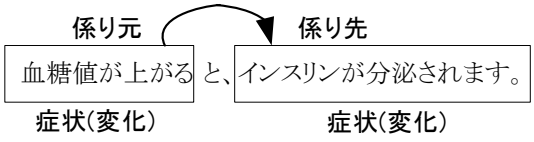

図 1. 係り受け関係にある節の例

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、健康番組において顕著に出現する以下の 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、表 1 に示すような 2x2 分割表を作成する。

表 1. 2x2 分割表

	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 を持つと判定できる。

4．健康に関する事柄を表現する節の曖昧性評価

3 章で述べた手法で関係を導き出した節の中には、必要な情報が省略され、どのような事柄を表現しているか曖昧なものが含まれる。例えば、「症状が出て、血腫が脳を圧迫する」という文からは、「症状が出る」という節が「血腫が脳を圧迫する」の原因として抽出される。しかし、「症状が出る」という病状の変化を表す節は、何の症状が出るかという情報が省略されているため、このままでは曖昧で知識として相応しくない。他の文の「肝硬変になり症状がでる」から「肝硬変になる」の症状として「症状がでる」という節が抽出された場合、「肝硬変になる」の症状として、「症状がでる」の原因となる「血腫が脳を圧迫する」が誤導出される危険もある。そこで、節の曖昧性を評価して、健康に関する事柄の表現として情報の不足があるかを判定する。曖昧性の評価を行い曖昧性が低いと判定された節間の関係のみを良質の知識として蓄積することができる。

曖昧性の高い節は、以下の条件を満たすという仮説

を立て評価する。

- 節を修飾する文節（時間表現や人物表現は除く）の種類数が多い
- 節を修飾する文節（時間表現や人物表現は除く）の出現頻度が一様に高い
- 節の直前に区切れ目（句点、読点、接続詞など）の出現が少ない

これらの条件を考慮するために、節の直前までに位置する語のエントロピーを用いて節 p の曖昧性を評価する値 $H(p)$を求める。コーパス中で節 p を修飾する文節中の自立語 x_i（名詞または形容詞で時間表現や人物表現は除く）の出現頻度を $N(x_i; p)$、節 p のコーパス中での全出現頻度を $C(p)$としたとき、節 p の曖昧性を評価する値 $H(p)$を以下の式で定義する。

$$H(p) = -\sum_{x_i} \frac{N(x_i; p)}{C(p)} \log \frac{N(x_i; p)}{C(p)}$$

この式では、節を修飾する文節の種類が多く、各頻度が一様であるほど、$H(p)$の値は大きくなり、また、節の直前に区切れ目の出現が多いほど、各自立語の出現確率（$N(x_i; p)/ C(p)$）が小さくなるため、$H(p)$の値は小さくなる。

5．実験

NHK で放送された「きょうの健康」の糖尿病に関連する番組を処理対象として、健康に関する事柄間の関係を推定する実験と、この処理において抽出された健康に関する事柄を表現する節の曖昧性を評価する実験を行った。以下に各実験の詳細を記す。

5.1 健康に関する事柄間の関係推定実験

「きょうの健康」80 番組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表 2 に示す。

表 2. 節の意味カテゴリー分類評価結果

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 項目を表 3 に示す。

表 3. 特徴的な節ペアと節間の関係抽出結果

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、テストデータから節間の関係を推定した。結果の一部を表 4 に示す。さらに、推定結果に対して、原因、症状、目的、対処法の 4 つの関係をまとめて評価した結果を表 5 に示す。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。

5.2 節の曖昧性評価実験

言語資源「Web 日本語 N グラム第 1 版」[4]を利用して、4 章で説明した節の曖昧性を評価する値 $H(p)$を計算した。この Web 日本語 N グラムには、Google 社が

表 4. テストデータから抽出した健康に関する事柄を表現する節とその関係（一部）

係り元の節	係り先の節	節間の関係
コレステロールが溜る[症状(変化)]	血管が狭窄する[症状(変化)]	原因
食事を摂る[行為(体の動作)]	血糖値が上昇する[症状(変化)]	原因
腎症が進む[病気(変化)]	高血圧が助長される[症状(変化)]	症状
腎症神経症を起こす[病気(変化)]	細い血管が変化する[症状(変化)]	症状
簡単に指先で血を採る[行為(医療)]	血糖を測る[行為(医療)]	目的
インスリンを注射する[行為(医療)]	血糖値を下げる[行為(医療)]	目的
合併症がある[病気(状態)]	栄養のバランスに気をつける[行為(管理)]	対処法
糖尿病にならない[病気(状態)]	食事を考える[行為(管理)]	対処法

表 5. 節間の関係評価結果

適合率	再現率
81.0% (51/63)	31.3% (51/163)

日本語の Web ページをクロールすることにより獲得した約 200 億文を対象として、出現頻度 20 回以上の 1～7 グラムの形態素列の情報が含まれている。例えば、節“症状が出る”に対して計算する場合、この節は 3 形態素で構成されているため 3 グラムのデータから「“症状”, “が”, “出る（他の活用形も含む）”」頻度により、節のコーパス中での全出現頻度が算出できる。また、4 グラム、5 グラムのデータから、節を修飾する文節の種類と頻度を抽出することができる。健康に関する事柄を表現する節に対する曖昧性評価結果の一部を表 6 に示す。

表 6. 節に対する曖昧性評価結果（一部）

$H(p)$	健康に関する節	$H(p)$	健康に関する節
4.79	機能が低下する	0.98	梗塞が増える
3.55	状態が続く	0.85	腫瘍ができる
3.03	反応が起こる	0.77	神経が緊張する
2.45	細胞が増える	0.65	緊張状態が続く
2.16	関節が痛い	0.54	やけどをする
2.08	症状が出る	0.42	風邪をひく
1.80	感染を起こす	0.42	中耳炎を起こす
1.79	神経に働く	0.28	血圧が高い
1.33	血管が詰まる	0.18	血糖値が上がる
1.20	粘膜が腫れる	0.13	脳出血を起こす

“機能が低下する”、“状態が続く”などの曖昧性が高いと考えられる節に対する $H(p)$ の値が大きく、曖昧性の低いと考えられる節に対しては小さな値が与えられている。この結果を評価するために、表 6 に示す 20 個の節に対して 2 人の被験者により曖昧の度合いを主観的に順位付けしてもらい、その平均による順位と、表 6 に示す $H(p)$ の値による順位との相関を計算した。また、比較対象手法として、4 章で言及した仮説の一つである、節の直前に出現する区切れ目の出現割合を基準とした手法を利用した。曖昧性が高い節ほど、その直前に修飾語が多くなると考えられるため、節の直前にある区切れ目の出現割合の昇順に曖昧性が高いと判定した。

順位の相関を計算する手法として、以下の式で示されるスピアマンの順位相関係数を用いた。

$$\rho = 1 - 6\sum D^2 / N(N^2 - 1)$$

ここで、D は 2 つの順位の差を、N は対象データペアの数（本実験では 20）を示す。ρ の値が 1 に近いほど 2 つのデータは強い正の相関があり、0 に近いときは相関が弱いと考えられる。逆に ρ の値が-1 に近い時は、2 つのデータには負の相関があると考えられる。順位の相関を計算した結果、$H(p)$の値による順位と、人手による順位では $\rho = 0.766$ 、節の直前に出現する区切れ目の出現割合を利用した手法と、人手による順位では $\rho = 0.461$ であった。提案手法による結果は、節の直前に出現する区切れ目だけを利用する手法に比べて、人手による順位との強い正の相関があり、この値は節の曖昧性評価のための一定の指標になり得ると判断できる。

6. 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した。また、健康に関する事柄を表現する節の曖昧性を評価する手法も提案した。実験により、健康に関する節間の関係を自動獲得でき、節の直前までに位置する語のエントロピーに基づく式により、節の曖昧性を評価できることを確認した。今後、節の曖昧性評価において、節の曖昧性の有無の判定のための閾値を設定することにより、自動獲得した節間の関係から知識として利用できるものの取捨選択を行い、クローズドキャプションからの知識獲得へと進めていく。さらに、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] 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)

[2] 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)

[3] William L. Hays, Statistics: Analyzing Qualitative Data, Rinehart and Winston, Inc., Chapter18, pp769-783 (1988)

[4] 工藤, 賀沢, Web 日本語 N グラム第 1 版、言語資源協会

[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)