E-001

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

A Study on Causality Extraction Targeting Closed-Captions

山田 一郎† 小早川 健† 三浦 菊佳† 住吉英樹† 八木伸行† 崔杞鮮‡
Ichiro Yamada Takeshi Kobayakawa Kikuka Miura Hideki Sumiyoshi Nobuyuki Yagi Kei-Sun Choi

1．はじめに

ディジタル放送では、データ放送や字幕放送など大量の信頼できるテキストデータが多重され放送されている。受信機が、このテキストデータを常時監視し、有益な情報を抽出、蓄積できれば、視聴者からの質問に何でも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、番組のクローズドキャプションを対象として、単語間における「原因－結果」の関係（以後、因果関係と呼ぶ）を自動抽出する研究を進めている。

乾ら[1]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 つの句が、いずれに属するかを推定する手法を提案している。鳥澤[2]は、並列句が一つの文に存在し、並列句中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。また、Chang ら[3]は、英文テキストを対象として、因果関係にある単語ペアと構文構造を学習する手法を提案している。この手法では、手掛かり語を特定しないでも因果関係にある単語ペアと構文構造が精度良く抽出でき、並列句中の動詞が共通の目的語を持たない場合にも対処可能であるため、汎用的な手法と考えられる。

本稿では、Chang らの手法をベースにして、日本語テキスト（TV 番組「きょうの健康」のクローズドキャプション）から因果関係のある名詞ペアと名詞ペア間の構文構造を抽出する手法を提案する。

2．因果関係表現の分類

テキスト中における因果関係の表現は、以下の 3 種類に分類できる。

(1) 同一文の句のペアに因果関係がある場合
「急に運動を始める(原因)と血圧が急上昇します(結果)。」
(2) 同一文の名詞ペアに因果関係がある場合
「脳卒中(結果)の原因となる動脈硬化(原因)が促進される。」
(3) 複数文にまたがって因果関係のある名詞ペア、句のペアが出現する場合
「心臓肥大(原因)が促進される。」
「この結果、心筋梗塞(結果)が起こりやすくなる。」

(1)は、2 つの句に含まれる名詞ペア("運動","血圧")を対象とすることにより、(2)に帰着できると考えられる。そこで本報告では(2)を対象とし、名詞ペアと名詞ペア間の構文構造に注目して、因果関係の有無を判定する。(3)は、接続詞などが手掛かりになるであろうが、難しい問題と考え、今回の対象外とした。

† NHK 放送技術研究所
‡ Korea Advanced Institute of Science and Technology

3．因果関係抽出処理

名詞ペアと名詞ペア間の構文構造に注目して 2 つの名詞間に因果関係が有るかを判定するために、Nigam らが提案した Naïve Bayes の分類器に EM アルゴリズムを組み合わせた手法[4]を利用する。Nigam らは、ラベル付き訓練データを利用してラベル無しデータのラベルを推定することにより、テキスト分類を行なっている。本手法では、少量のクローズドキャプション中の名詞ペアに、因果関係の有無を判定したラベルを付与し、ラベル無しのクローズドキャプションのラベルを推定する。以下にその処理概要を記す。

3.1 テキストからの特徴抽出

まず、入力されるクローズドキャプションテキストから名詞ペアと名詞ペア間の構文構造を抽出する。対象を分類語彙表[5]により因果関係を表現しやすい名詞に限定し、南瓜[6]による構文解析結果を利用して名詞ペア間がどのような構文構造に位置しているかを抽出する。そして、Preorder String Expression[7]（以後、PSE と呼ぶ）により構文構造の表現とマッチング処理を行なう。ここでは、南瓜により分割された句を自立語と機能語の 2 つに分割して PSE の構造に利用した。以下に、PSE の例を示す。

[入力 1] 名詞 1 が起きると名詞 2 につながります。
[PSE1] {"つながる","と","起きる","が", "名詞 1", 0,0,0,0," に","名詞 2",0,0,0}
[入力 2] 名詞 1 が名詞 2 につながる
[PSE2] {"つながる","が", "名詞 1", 0,0,"に","名詞 2",0,0,0}

PSE の表現では、語順と要素"0"により元の構文構造を復元することができる。2 つの PSE p_1, p_2 に出現する名詞ペア間の構文構造の類似性 $sim(p_1, p_2)$を以下の式により評価する。

$$sim(p_1, p_2) = \frac{com(p_1, p_2) \times 2}{wc(p_1) + wc(p_2)}$$

ここで、$wc(p_i)$は PSE p_iに出現する名詞 1 と名詞 2、0 以外の単語数、$com(p_1, p_2)$はそのうちの PSE の構造を考慮した共通単語数を示す。例えば上記の例では、単語"つながる"、"が"、"に"が共通単語になるため、sim(PSE1, PSE2) = 6/8 = 0.75 となる。この類似性評価の値は EM アルゴリズムで利用する。

テキストから抽出した 2 つの名詞は、分類語彙表上での属性が一意に決まる場合はその 5 桁目までの数値を、複数の属性を持つ場合は表記そのものを利用する。2 つの名詞と、その間の構文構造と合わせて 3 項組として扱う。例えば、2 章(2)の例は以下のように表現される。

<15721, 15721, {名詞 2, "なる","と","原因","の",名詞 1}>

この 3 項組が因果関係を表すか否かを評価する。

3.2 Naïve Bayes

抽出された 3 項組 t_i が因果関係を持つ(c_1)、もしくは持たない(c_0)確率は、以下の式で与えられる。

$$P(c_j \mid t_i) = \frac{P(c_j)P(t_i \mid c_j)}{P(t_i)}$$

この値が大きいクラス c_j(c_0 または c_1)を、因果関係の有無の判定結果とする。$P(t_i \mid c_j)$は、以下の式とする。

$$P(t_i \mid c_j) = P(CP_{t_i} \mid c_j)P(SP_{t_i} \mid c_j)$$

ここで、CP_{ti} は 3 項組 t_i に含まれる 2 つの名詞間の構文構造を指し、SP_{ti} は 3 項組 t_i に含まれる名詞ペアを指す。この式を利用して、ＥＭアルゴリズムにより $P(c_j \mid t_i)$を推定する。

3.3 EM アルゴリズム

EM アルゴリズムは、内部状態が不明な不完全データに対して尤度が最大になるような繰り返し学習を行ない、内部状態を推定する手法であり、この場合は教師無しデータが不完全データとなる。まず、すべてのクローズドキャプション集合を対象として、あるクラス c_j のもとで素性となる CP_{ti}、SP_{ti} が発生する確率 $P(CP_{ti} \mid c_j)$、$P(SP_{ti} \mid c_j)$を以下の式により求める(M ステップ)。

$$P(CP_{t_i} \mid c_j) = \frac{1 + \sum_{k=1}^{|T|} sim'(CP_{t_i}, CP_{t_k})P(c_j \mid t_k)}{|CP| + \sum_{m=1}^{|CP|}\sum_{k=1}^{|T|} sim'(CP_{t_m}, CP_{t_k})P(c_j \mid t_k)}$$

$$P(SP_{t_i} \mid c_j) = \frac{1 + \sum_{k=1}^{|T|} N(SP_{t_i}, t_k)P(c_j \mid t_k)}{|SP| + \sum_{m=1}^{|SP|}\sum_{k=1}^{|T|} N(SP_{t_m}, t_k)P(c_j \mid t_k)}$$

ここで、$|CP|$、$|SP|$、$|T|$は、名詞間の構文構造の総数、名詞ペアの総数、3 項組の総数を表し、$N(SP,t_k)$は 3 項組 t_k に名詞ペアが含まれるか否かを表す関数であり、含まれるときだけ 1 の値を取る。$sim'(CP_{ti}, CP_{tk})$は名詞ペア間の構文構造の類似性で、0.5 より大きい場合に $sim(CP_{ti}, CP_{tk})$を、それ以外は 0 を与える。$P(c_j \mid t_k)$の初期値は、因果関係の有無を判定した少量のクローズドキャプション（教師有り訓練データ）を利用して計算する。

次に、Naïve Bayes の式を利用して、$P(c_j \mid t_i)$の期待値を計算する（E ステップ）。

$$P(c_j \mid t_i) = \frac{P(c_j)P(CP_{t_i} \mid c_j)P(SP_{t_i} \mid c_j)}{\sum_r P(c_r)P(CP_{t_i} \mid c_r)P(SP_{t_i} \mid c_r)}$$

$$P(c_j) = \frac{1 + \sum_{k=1}^{|T|} P(c_j \mid t_k)}{|c| + |T|}$$

$|c|$は分類すべきクラスの数を指し、ここでは 2 となる。M ステップと E ステップを繰り返すことにより、クローズドキャプションに出現する 3 項組が因果関係を持つか否かを $P(c_j \mid t_i)$の値により推定できる。さらには、$P(CP_{ti} \mid c_j)$からベイズの定理により $P(c_j \mid CP_{ti})$が計算可能で、因果関係を持つ時の特徴的な構文構造の判定が可能となる。

4．実験

前章までの手法の検証のために、循環器系の話題を取り上げている「きょうの健康」16 番組を対象とし、番組で使われたクローズドキャプション 2180 文から 3 項組 1495 個を抽出して因果関係抽出実験を行なった。無作為に 1 番組を選び、そこから抽出した 3 項組 149 個に対して人手により因果関係の有無をタグ付けして教師有り訓練データとし、残りの 15 番組を教師無しデータとした。繰り返し回数を $P(c_j)$の収束度合を基準として判定したとき（実験では 100 回）、因果関係を持つと判定された 3 項組を生成する原文（一部）を表 1 に示す。

表 1. 抽出された因果関係（一部）

$P(c_1 \mid t_i)$	3 項組を生成する原文（括弧内が対象名詞）
0.980	[コレステロール]が高いほど[心筋梗塞]の危険も高くなる
0.978	[コレステロール]が高いと[冠動脈疾患]が起こりやすい
0.973	[動脈硬化]が進んで[虚血性心疾患]を起こす
0.962	[中性脂肪]が高いのは[動脈硬化]の危険信号、と考える
0.955	[脳卒中]の原因となる[動脈硬化]が・・・
0.931	[動脈硬化]を起こすリスク例えば[高血圧]を持つ

表 1 の名詞ペアとその間の構文構造は、いずれも訓練データには存在せず、EM アルゴリズムにより獲得できたものである。実験で利用した教師無しデータ中の 1 番組を取り出し適合率を評価した結果、因果関係がある名詞ペアは 73.8%(31/42)、無い名詞ペアは 61.9%(75/121)であった。因果関係が無い名詞ペアに対する結果が悪い。因果関係が無い時の名詞ペアや名詞ペア間の構文構造のパターンは無数に存在するため、少量の訓練データのみの学習では不十分であったと考えられる。今後、データ量を増やした実験を行なうこととしたい。

5. まとめ

本稿では、因果関係の有無を判定した少量の日本語テキストを利用して、ラベル無しの日本語テキストから因果関係を持つ 3 項組を抽出する手法を提案した。TV 番組「きょうの健康」を対象とした実験により、因果関係の候補が獲得できることを確認した。

EM アルゴリズムではループ回数により結果が異なることがある。新納らは、データを分割して交差検定を行なうことにより、その繰り返し回数を推定する手法を提案している[8]。今後、このような技術を取り入れ、データ量を増やした実験を通して精度の向上をはかる予定である。

【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] Du-Seong Chang, Key-Sun Choi: Causal Relation Extraction Using Cue Phrase and Lexical Pair Probabilities. *IJCNLP 2004*, pp.61-70(2004)

[4] Kamel Nigam et al.: Text Classification from Labeled and Unlabeled Document using EM. *Machine Learning*, Vol.39, No.2/3, pp.103-134(2000)

[5] 国立国語研究所: 分類語彙表 増補改訂版(2004)

[6] 工藤ほか: チャンキングの段階適用による係り受け解析, 情処論, Vol.43, No.6, pp.1834-1842(2002)

[7] Fabrizio Luccio et al.: Exact Rooted Subtree Matching in Sublinear Time, Technical Report TR-01-14(2001)

[8] 新納ほか: EM アルゴリズムの最適ループ回数の予測を用いた語義判別規則の教師無し学習, 情処論, Vol.44, No.12, pp.3211-3220 (2003).