E-013

半教師あり学習による事物間の関係を表現する典型的な構文構造の抽出

A Study on Extraction of Typical Sentence Structures Expressing Object Relations using a Semi-supervised Learning

山田 一郎† 三浦 菊佳† 小早川健† 住吉英樹† 柴田正啓† 八木伸行†
Ichiro Yamada Kikuka Miura Takeshi Kobayakawa Hideki Sumiyoshi Masahiro Shibata Nobuyuki Yagi

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、放送番組を利用した様々なアプリケーションを実現できる。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、番組で発生する事象間や登場する事物間の関係を特定する研究を進めている。

テキストデータから関係を抽出する従来研究として、乾らは、「ため」という単語を手掛かり語として取り出した 2 つの節間の関係を推定する手法を提案している[1]。この手法では、関係として“ 原因”、“ 効果”、“ 前提条件”、“ 手段”という 4 種類を対象としている。また、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]。我々は、「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を抽出する手法を提案している[3]。これらの手法では、あらかじめ対象とする“ 関係”を決めてから、その関係にあるテキスト中の用語や節の抽出を試みている。しかし、この関係自体もテキスト中に記述されていることがある。例えば以下の文では、“ マングース”と“ ヘビ”の間に“ 天敵”という関係があることが明示されている。

「マングースはヘビの天敵です。」

また、この文で使われている表現「～は～の～です。」は、他にも例えば「レアはダチョウの仲間です。」など頻出しており、事物間の関係を表現する典型的な構文構造と考えられる。そこで本稿では、あらかじめ対象とする関係を決めることなく事物間の関係を特定し、さらに、事物間の関係を表現する典型的な構文構造を抽出する手法を提案する。NHK で放送された「地球！ふしぎ大自然」という番組のクローズドキャプションを対象とした動物間の関係抽出処理と実験について報告する。

2．関係推定処理

事物間の関係を特定し、事物間の関係を表現する典型的な構文構造を抽出するために、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半教師有り学習手法[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事物ペアに、関係有ラベルを付与し、ラベル無し（関係の有無が不明）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。

2.1 テキストからの特徴抽出

事物間の関係を推定する対象を“ 動物”とし、まずはクローズドキャプションの同一文に出現する動物ペアを抽出する。動物の抽出処理では、手作業で生成した動物名辞書を利用する。次に、動物ペアの共通係り先となる文節を特定し、各動物から共通係り先までの係り受け構造と、この係り受け構造に含まれずに共通係り先の文節を修飾する係り受け構造を構文構造特徴として抽出する。最後に、2 つの動物間の関係候補となる単語を構文構造特徴から抽出する。この関係候補となる単語は、名詞以外に形容詞、動詞なども対象とする。例えば以下の文から特徴を抽出する。

「プレーリードッグにとってイヌワシは恐ろしい天敵です。」

動物ペア：「プレーリードッグ」「イヌワシ」
共通係り先文節：「天敵です。」
各動物から共通係り先までの係り受け構造：
　動物 1（プレーリードッグ）：「にとって」
　動物 2（イヌワシ）：「は」
共通係り先を修飾する係り受け構造：「恐ろしい」
関係候補：「恐ろしい」「天敵」

この場合、関係候補として「恐ろしい」と「天敵」の 2 つが抽出されるため、最終的に以下の 2 種類の特徴が生成される。

プレーリードッグ|イヌワシ|恐ろしい|NP1,にとって=NP2,は=REL,NULL,天敵
プレーリードッグ|イヌワシ|天敵|NP1,にとって=NP2,は=恐ろしい,NULL,REL

この特徴では、セパレータ“ |”で分割された第一項目と第二項目が動物ペア、第三項目が関係候補、第四項目が構文構造特徴となる。構文構造特徴では、動物名と関係候補は抽象化し(NP1、NP2、REL)、各動物から共通係り先までの係り受け構造と、共通係り先を修飾する係り受け構造を、セパレータ“ =”で区切っている。また、各文節は自立語と付属語に分けて扱い、付属語が存在しない文節には NULL とする。この特徴を半教師有り学習手法の入力とする。

2.2 半教師有り学習手法

前節で抽出された特徴は、動物ペア、関係候補、構文構造により構成される。この 3 種類の項目が、どの程度、関係を示す表現として用いられるかを確率値として定式化し、各確率値を推定することにより、3 種類の項目全体が関係を表す確率値を推定する。

抽出された 3 種類の項目全体 t_i が関係を持つ(c_1)、もしく

† NHK 放送技術研究所

は持たない(c_0)確率を、以下の式で与える。

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

この値が大きいクラス c_j(c_0 または c_1)を、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。

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

ここで、SP_{ti}は t_iに含まれる動物ペアを、RP_{ti}は t_iに含まれる関係候補を、CP_{ti} は t_i に含まれる構文構造を指す。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。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 SP_{ti}、RP_{ti}、CP_{ti} が発生する確率 $P(SP_{ti}|c_j)$、$P(RP_{ti}|c_j)$、$P(CP_{ti}|c_j)$を以下の式により求める(M ステップ)。

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

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

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

ここで、$|SP|$、$|RP|$、$|CP|$、$|T|$は、動物ペアの総数、関係候補の総数、構文構造の総数、3 種類の項目全体の総数を表し、$N(SP,t_k)$は 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。$P(c_j|t_k)$の初期値は、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、$P(c_j \mid t_i)$の期待値を計算する（E ステップ）。

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

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

$|c|$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 3 種類の項目が出現したときに、その表現が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、$P(SP_{ti}|c_j)$、$P(RP_{ti}|c_j)$、$P(CP_{ti}|c_j)$からベイズの定理により $P(c_j|SP_{ti})$、$P(c_j|RP_{ti})$、$P(c_j|CP_{ti})$が計算可能で、関係を持つ動物ペア、関係を現す単語、関係を表す時の特徴的な構文構造の判定が可能となる。

3．実験

前章までの手法の検証のため、NHK で放送された「地球！ふしぎ大自然」116 番組を対象とした関係を表す時の特徴的な構文構造を推定する実験を行った。番組で使われたクローズドキャプションから複数の動物が同一文中に出現するテキストを抽出し、3170 組の動物ペア、関係候補、構文構造で構成される特徴を生成した。この中で、関係候補に“天敵”、“仲間”、“弱い”、“食べる”がある特徴 101 組は、明らかに関係を表すと判断して $P(c_1|t_k)$の初期値を 1 とし、$P(c_j)$の変化が十分に小さくなる繰り返し数 100 回で実験を行った。構文構造が関係を表す確率 $P(c_1|CP_{ti})$の計算結果の一部を表 1 に示す。

表 1. 構文構造が関係を表す確率計算結果例
(NP1,NP2:動物ペア　REL:関係候補)

$P(c_1\|CP_{ti})$	構文構造
0.964	NP1,は=NP2,を=REL
0.954	NP1,が=NP2,を=REL
0.952	NP1,は=NP2,の=REL
0.902	NP1,の,REL,NULL==NP2
0.883	NP1,を,REL,NULL==NP2
0.826	NP1,にとって=NP2,は=REL
～	
0.293	NP1,や,REL, NULL ==NP2
0.288	NP1,と=NP2,の=REL
0.081	NP1,NULL,REL,NULL==NP2

1 章の例で挙げた「マングースはヘビの天敵です。」の構文構造(NP1,は=NP2,の=REL)は $P(c_1|t_k)$=0.952 と高確率であり、一方、3 つの名詞が並列で出現する構文構造(NP1, NULL,REL,NULL==NP2)では、その確率値が低い。この結果から、関係を表す構文構造と、関係を表さない構文構造とを弁別できていることがわかる。

4．まとめ

本稿では、半教師有り学習を利用することにより、番組のクローズドキャプションから関係を表す時の特徴的な構文構造を抽出する手法を提案した。実験により、動物の関係を表す時の構文特徴と関係を現す単語、関係する動物ペアを抽出できることを確認した。

提案手法の処理結果を利用することにより、放送した番組映像を利用して様々な動物に関する調べ学習を可能とするマルチメディア百科事典[5]や、視聴中の番組内容に関連する情報を自動提示するシステム CurioView[6]の一機能を実現することができる。

実験では動物を処理対象としたが、他の事物に対しても同様のアルゴリズムで処理可能であると考えられる。今後、他の事物を対象とした実験を進める予定である。

【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] 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)
[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] 三浦ほか:放送番組を素材としたマルチメディア百科事典の自動構築, 映像情報メディア学会誌, Vol.62, No.1, pp.110-116(2008)
[6] 住吉ほか: CurioView：情報検索を活用した新しい視聴スタイルの提案, 映像情報メディア学会年次大会(2008)