1V-8

イベント共参照関係を利用した因果関係知識の獲得

田中 翔平 † 岡崎 直観 ‡ 石塚 満 †
† 東京大学大学院情報理工学系研究科 ‡ 東北大学大学院情報科学研究科 † 東京大学大学院情報理工学系研究科

1 はじめに

近年，情報抽出の研究が盛んに行われ，2つのエンティティとそれを結び付ける関係，という知識を獲得するために様々な手法が提案されてきた．しかしながら，膨大な数を抽出することが可能になったそのような知識を，更に結び付けるのような知識を獲得することは非常に困難である．本研究ではこの問題に着目し，イベント間の共参照関係を手掛かりとして，関係が導く知識，及び複数の関係から導かれる知識を獲得することを試みる．

2 手法

2.1 システム概要

本研究での最終的な出力は，単数あるいは複数のイベント(関係) が成立するとき，それによって導かれる知識 (そのイベントが原因で起こりうることなど) である．本研究では入力として，最初に1つの関係を表す動詞を与える．その動詞から導かれる知識を探すために，その動詞が記述するイベントを参照する表現として，名詞化表現を用いる．本研究では，そのようなイベントの共参照表現が出現する文を深く解析することで，入力動詞が表現するイベント (関係) によって導かれる知識をパターンとして獲得する．更に，その知識が成立するために必要な別の関係を求めることで，複数の関係から導かれる知識を獲得する．

2.2 名詞化表現の獲得

本研究では，関係によって導かれる知識を獲得するためのイベント共参照表現として，動詞の名詞化表現を利用する．動詞の名詞化表現は，Flamenet[1]を利用して獲得する．Flamenetには様々な意味的な格 (フレーム) と，そこに属する単語が掲載されており，意味的に近い単語のグループを獲得することができる．例えば，動詞”acquire”はフレーム”Getting”に属しており，フレーム”Getting”には他に”get”や”obtain”などが属する．フレーム”Getting”には名詞である”acquisition”も属しており，これは”acquire”を受ける名詞と見なすことができる．本研究では，このように与えた動詞と同じフレームに属する名詞を全て，動詞が記述するイベントを受ける表現として抽出する．また同時に，そのフレームに属する動詞も全て，与えた動詞と同じ意味を表現するものとして抽出する．

2.3 知識抽出対象となる文書の獲得

次に，知識抽出に用いる文書を獲得する．Flamenet から獲得した動詞と名詞化表現を利用して，コーパスに対してそれらが共起する文書を検索する．この時，動詞，名詞化表現の中で，それぞれいずれか1つでも文書内に出現すれば，知識抽出対象とする．つまり，例えば動詞”acquire”と名詞”acquisition”が共起する文書だけでなく，動詞”get”と名詞”acquisition”が共起した文書も獲得する．文書を獲得したら，それぞれの文書に対し，品詞・構文解析，固有表現抽出，共参照解決 (この場合の共参照解決は代名詞等の解決である) を行う．各処理は全て，Stanford CoreNLP[2]を用いる．

2.4 パターンの作成

ここでは，関係によって導かれる知識を表現するパターンを作成する．最初に，入力動詞が表現するイベントの知識を抽出する．すなわち，入力動詞”acquire”に対し,「誰が」「何を」”acquire”したのか，という知識の抽出を行う．知識の獲得は，文書内で入力動詞 (及び Flamenet で拡張した動詞) を含む文を探し，構文解析上でその動詞の主語，目的語を抽出することで行う．固有表現抽出及び共参照解決の結果，その主語と目的語が固有表現でない場合は，知識を抽出しない．そうでない場合は，例えば”A acquire B”(A がある会社，B が別の会社) のような知識が獲得できる．なお，文書内でそのような知識が1つも抽出できない場合は，その文書を以後の処理から除外する．

次に，獲得した動詞によるイベントを受ける名詞化表現を含む文から，パターンを作成する．本研究では動詞が表現するイベント (関係) が導く知識を獲得することを目的とするため，名詞化表現が主語になるようなパターンを探す．すなわち，文書内で Flamenet から獲得した名詞のいずれかを含む文に対し，その名詞が主語になる構文解析上の部分木を抽出する．その部分木において，根 (root) となる動詞を探す．root となる動詞を中心として，再利用可能な知識を表現するパターンを作成するため，その部分木の抽象化を行う．まず root から見て，構文木上深さ 4 以下の単語はパターンに不必要と見なし，除外する．次に，深さ 3 の単語はパターン上の置き換え可能な変数と見なし，それを表現する”X”などに置き換える．ここで，”X”に入る単語 (X が名詞句の一部ならその全体) は別途保持しておく．深さ 2，1 の単語はそのままパターンとして利用する．残った単語の中で，固有表現が存在する場合は，固有表現の種類に従って変数で置き換える．例えば残った単語の中で，A や B(入力動詞の表現するイベントの主語・目的語) と見なすことができるものがあれば，”A”や”B”で置き換える．また，固有表現抽出で数字や金額等を示す表現であることが分かっている場合には，”NUMBER”や”MONAY”などの特殊な変数で置き換える．特殊な固有表現でも，A でも B でもない固有表現については，変数として”X”などで置き換える．深さ 3 の単語で置き換えたものの中で，固有表現に相当するものがあれば，それぞれ特殊な変数で置き換える．こうして変数を含む部分木が獲得できるが，root との距離が近くても，例えば同格表現や副詞による修飾，”and”表現など構文上不要な単語も存在するため，そのような単語は除去する．また，この木の主語である名詞化表現の修飾語も，不必要として除去する．最終的に，”X”に入る表現は名詞のみとなる．こうして，入力動詞が表現するイベントによって導かれる知識を表現するパターンが獲得できる．

2.5 関係の獲得

パターン上の変数と，入力動詞によって記述されるイベントを結び付けるため，イベントの主語である A，または目的語である B との関係を求める．関係を探すために，リソースとして Reverb[3]を ClueWeb09[4]に対して適応したものを用いる．このリソースには，(entity1, relation, entity2) の形での膨大な数の関係の知識が含まれる．A 及び B と，保持しておいた各変数に元々入っていた名詞 (及び名詞句) を用いて，両者を entity1 あるいは entity2 として所持する関係を探す．複数

Causal Relation Extraction using Event Coreference
†Shohei TANAKA, Mitsuru ISHIZUKA, Graduate School of Information Science and Technology, University of Tokyo
‡Naoaki OKAZAKI, Graduate School of Information Sciences, Tohoku University

[1]https://framenet.icsi.berkeley.edu/fndrupal/
[2]http://nlp.stanford.edu/software/corenlp.shtml
[3]http://reverb.cs.washington.edu/
[4]http://lemurproject.org/clueweb09.php/

表 1: ルールの評価

動詞	#rule	AT	FR	FP
acquire	69	29	15	25
visit	68	22	14	32
attack	60	23	6	30

見つかった場合には，Reverb のスコア，及び ClueWeb09 内での抽出頻度を利用して，一番上位のものを取得する．もし関係が見つからない場合は，各名詞句の一部などを用い，条件を緩めた形で検索する．それでも見つからない場合は，そのパターンは出力から除外する．

以上の一連の処理によって，例えば”A acquire B”と”B *relation* X”が成り立つ時，X を含むパターン (例えば”The acquisition give A an advantage in X”) が成り立つといったような，複数の関係を結び付ける知識を獲得することができる．

3 実験

3.1 内容

提案手法を実際にコーパスを用いていくつかの動詞に対して適用し，知識を抽出する実験を行う．コーパスとして，今回の実験では English Gigaword Corpus Third Edition(LDC2007T07; EGC) の中で，特に Los Angeles Times/Washington Post Newswire Service のニュース記事を用いた．入力として与える動詞は，EGC に出現する全ての動詞を頻度順に並べ，上位 200 個の中で，固有名詞 A，B を用いて”A *verb* B”の形で記述でき，かつ名詞化表現が明確である”acquire”(Getting)，”visit”(Visiting)，”attack”(Attack) の 3 つとした (括弧の中は Framenet でのフレーム)．本研究で扱うような知識に対して，良く知られた評価用データセットが存在しないため，ルールの評価は全て人手で行う．ただし，ルールの正誤だけでなく，ルールが誤りである場合パターンに誤りがあるのか，変数と A や B との関係に誤りがあるのか，という点まで評価を行う．

3.2 結果

結果を表 1 に示す．表中”AT”はルールが正しい場合を示し，”FR”はパターンは正しいが変数と A 及び B との関係に誤りがある場合，”FP”はパターンが誤りである場合を示す．ルールの抽出精度は”acquire”で 42.02%，”visit”で 32.35%，”attack”で 38.33%であった．それぞれの動詞について正しく抽出されたルールの例を，表 2 に示す．

表 2: ルールの例

”A acquire B” AND ”B have be fixate on X” → ”The acquisition will give A an advantage in X”
”A visit B” → ”The visit will spotlight A”
”A attack B” → ”The attack kill NUMBER person”

4 エラー分析

実験の結果，各動詞について表 1 に示すように知識として有用なものは多数獲得できたが，一方で多数のルールに誤りが生じた．まず，表 1 中の”FR”に相当する，パターンは正しいが変数と A や B との関係抽出において生じた誤りがある．その理由として，変数に元々入っていた名詞が，例えば”market”など，そのまま A と B との関係を求めても有用な関係が抽出できない場合や，ClueWeb09 で最頻出の関係が X と A や B との関係として相応しくなかった点が考えられる．これを解決するためには，今用いている簡易な関係の決定手法ではなく，より詳しい関係の評価手法が必要である．また，表 1 中の”FP”に相当するパターンの誤りであるが，これについてはいくつか別々の原因が存在している．まず，名詞化表現が主語になる部分木が，入力動詞の説明を行っている場合に生じる誤りがある．これは例えば，”The acquisition was held Friday, April 7, 2006...”のように，入力動詞が表現するイベントの付加的な情報を表現している場合である．次に，Flamenet を利用した名詞化表現の抽出で，本研究が対象とする知識を獲得するのに相応しくない名詞化表現を抽出したことによる誤りがある．例えば，動詞”visit”において，フレーム”Visiting”には名詞”visitor”が属するが，これは”visit”によって導かれる知識の獲得には相応しくない．これは，いくつか簡単なルールを作成することで，除去することが期待できる．更に，当然今回提案手法が利用したルールによる誤りも存在するが，この点については，更にルールを精査することで対応したい．最後に，そもそも係り受け解析や，自然言語処理において非常に難しい問題である代名詞等の共参照表現の解決に失敗したことによるエラーも多数存在する．これについては本研究では直接対処することはできないが，ある程度エラーが起き得る箇所については想定することができるため，構文解析を行う前の処理で対応する必要がある．

5 関連研究

本研究に関連する研究として，Do らの研究 [1] がある．この研究では文書内から，動詞が表現するイベントと，いくつかのルールを利用して作成した名詞化表現が記述するイベントを全て抽出し，それぞれのイベント間に因果関係が存在するかを判定することで，因果関係にあるイベントのペアを抽出している．本研究とは最終的に獲得したい知識も異なるが，特に名詞化表現の扱いが異なる．この研究では名詞化表現は動詞と同等のイベントを記述するものとして扱い，本研究のように動詞を受けるものとしては扱っていない．すなわち，例えば”A’s attack of B destroyed ...”のような表現が出現したとき，本研究では”destroy”以降に特に注目するが，Do らの研究では”A’s attack of B”を”A attack B”というイベントとして抽出し，同様にして抽出された他のイベントとの因果関係を判定している．

6 結論

本稿では，単数，及び複数の関係 (イベント) から導かれる知識を獲得するため，イベント共参照表現としての動詞の名詞化表現を手掛かりとして利用する手法を紹介した．また，獲得した知識を実際のコーパスから抽出し，人手による評価を行った．

今後は，エラー分析の結果を考慮した抽出精度の向上と，より多くの動詞からの知識の獲得，及びより定量的な評価を課題としたい．

参考文献

[1] Quang Xuan Do, Yee Seng Chan and Dan Roth, Minimally Supervised Event Causality Identification In *Proc. of EMNLP*, pages 294-303, 2011