表層的因果知識ベースによる事象推移予測方式

4Q－9

佐藤浩史　笠原　要　松澤和光
NTT コミュニケーション科学研究所

1．はじめに

我々は，人間が不完全な知識の下で行う概括的な判断（俗にいう「アバウトな判断」）に着目し，これを計算機で実現する処理方式を「アバウト推論」と名付けて研究を進めている [1]．この方式は，知識が欠落していても類似した常識の補完によって推論を進めることを特徴とする．

この推論に必要な常識ベースの一つとして，既に単語知識ベースである「概念ベース」を構築し，単語の意味を表す概念間の類似性判別方式を提案した [2]．そして次に「何をどうすればどうなるか」といった常識的な現象，すなわち事象間の因果関係を知識ベース化し，その推移を予測する一連の方式の研究を進めている．

既存の事象予測方式としてエキスパートシステムに用いられる方式があるが，これはドメインを限定した場合のものであり，我々が望む常識的な推論には向いていない．また，大規模な常識知識の構築プロジェクトである Cyc [3] は，人手により構築するため時間・コストがかかるという問題がある．

そこで，テキストコーパスを知識源とする「表層的因果知識ベース(Simplified Causality Base, SCB)」の構築法の提案を行う．この提案方法は因果知識の構造を簡略化し，構文解析ツールを用いることで自然言語テキストからの自動知識獲得を目指した物である．本稿では実際に知識獲得の実験を行い，検証する．

2．表層的因果知識

一般に自然言語テキストが持つ因果知識は，表面上は単なる原因と結果の関係であっても，実際は深い根拠がある場合が多い．

例えば，

「夏になると，半袖シャツの人が増える．」

という文は直接的には

「（季節が）夏になる」
→「半袖シャツの人が増える」　（＊）

という因果と考えられるが，実際は例えば，

「（季節が）夏になる」→「気温が上がる」
→「暑い」→「不快」
「半袖シャツを着る」→「体温を逃がしやすい」
→「涼しい」→「快適」

等の過程を経て，結論を得るとも言える．ところがこういった深い知識はきりがなく，漏れなく正確に記述することは困難であり，計算機が読み込み可能なメディアによる知識も少ない．これは，「当たり前のことはわざわざ記述されない，記録されない」という理由による．

そこで我々は深い知識を全て獲得するのではなく，上の例の(＊)の様に自然文テキスト上に直接存在する「表層的因果知識」(Simplified Causality,SC)に着目する．SC を大量に蓄えれば，深い因果は理解していなくても，人間が行う推移予測をある程度再現できると考えられる．このコンセプトは，「門前の小僧，習わぬ経を読む」という諺によく表されている．

2.1 表現形式

SC の原因および結果にあたる各事象を表現するためには，動作（用言）を中心とし，その動作対象との関連を体系的に整理する必要がある．そこで我々は，事象を表現するのに自然言語の格フレームを用いる．さらにテキストコーパスが持つ知識のスパースネスを解消するために，格フレームの動作対象となる体言をシソーラスを用いて一般化する．これらの代表的な物として，日英自動翻訳システム ALT-J/E の結合価パターン（約 15,000種）（図1），および ALTシソーラス（約 3,000 カテゴリ）がある [4]．

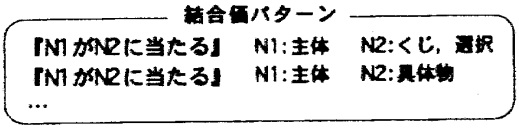

図1　格フレームの例

因果知識は，因果関係を持つ事象同士を連結したものであるから，各事象をノードとしたグラフ構造（ネットワーク）で表現できる．すなわち，原因となる事象から結果となる事象を有向アークで結び，そこに因果の形態情報および度合をラベルとして添付する．これを SC の1ユニットとする（図2）．

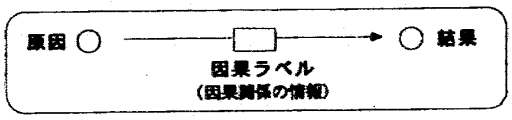

図2　SC のユニット

2.2 テキストからの知識自動獲得

自然言語テキストからの因果知識 (SC) 自動獲得法について説明する．

まず与えられたテキスト中の複文・重文を抽出する．これらの文は単文が接続された物なので，因果を含んでいるとみなすことができる．次に，構文解析ツールを用いて複文を格フレームに分解し，それぞれの格フレーム間の接続関係を獲得し，この接続の種類を因果の形態とする．

Transition Inferring with Simplified Causality Base
Hiroshi SATO,Kaname KASAHARA and Kazumitsu MATSUZAWA
NTT Communication Science Laboratories
email : hiroshi@cslab.kecl.ntt.co.jp

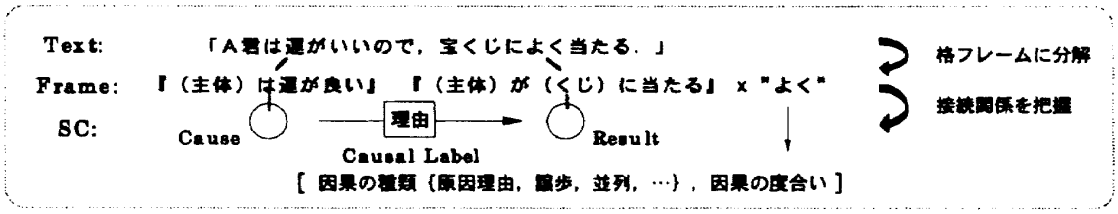

図3 因果知識獲得の構成

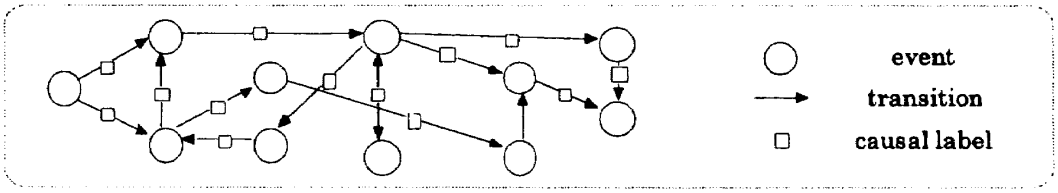

図4 因果知識グラフ

その際に，1文中の用言を強調する副詞，及び同一視するSCの全文を通しての出現頻度により因果の度合を決定し，ラベルに加える（図3）．

これらのSCユニット群で構成されるのが，因果知識グラフ(SC Graph)である（図4）．

実際に事象推移予測を行う際は，ラベルの情報を元に推移先の事象を選択する．

3．実験

上述の手法により因果知識獲得実験を行った．実験には，ALT-J/Eの結合価パターン，ALTシソーラス，ALT-J/Eの構文意味解析ツールを用いた．ソースとなるテキストには，常識的でなおかつ比較的単純な文が望ましいことから，EDRコーパス[5]の中から用例集を出典とする文を選んだ（約17,000文）．以下が実験結果の統計である．

有効文数 / 総文数	8,417/16,946
出現フレーム数	4,944
出現フレーム延べ数	16,834
平均次数	3.4
最大次数	221

図5 SC統計結果

ここで有効文とは，1文中に2つ以上の格フレームを持ち，かつ接続関係が認められる文である．今回の実験では入れ子状の接続関係，すなわち接続関係にある2つの文中にさらに多重の接続関係が認められた場合は，最も外側の接続関係のみを抽出しているので，それぞれ1つの有効文から1つずつの因果関係を獲得している．従って，出現フレームの延べ数がちょうど有効文数の2倍となる．また，次数とは各ノードに接続されるアークの本数である．

考察として，平均次数が3を超えていることから，因果ユニット同士が離散的にならず，ある程度接続していると言える．テキスト中の知識量についても，全文の約半分が有効であり，さらに入れ子状の接続関係の獲得を行えば十分な量が獲得できると思われる．

今回獲得した因果の中から，同一の原因事象を持つ複数の因果を抽出し，その一部の例を以下に挙げる（図6）．ここでは事象のキーとなる体言も考慮に入れ，格フレーム及び体言が一致したものを同一事象とみなしている．また，簡単のため格フレームは用言のみを表示している．

原因事象		結果事象	
用言	体言	用言	体言
見る	映画	変わる	イメージ
		残る	感動，胸
発生する	事件	直行する	パトカー
		逮捕する(受動)	犯人
成る	冬	渡る	鳥，日本
		絶える	村，連絡
		変わる	地面，雪

図6 獲得SC例

今回の実験では，副詞や副詞節の考慮を行っていないものの，人間の感覚に矛盾しない結果が得られている．この結果より，規模を拡大していけば本提案方式によりさらに質の良い知識ベースが構築できると期待される．

4．まとめ

事象推移予測システム実現のため，自然言語からの事象間の因果知識の自動獲得法の提案を行った．今後は知識ベースの規模拡大とともに，事象推移予測を行う方式の検討を行う予定である．

参考文献

1. 松澤，石川，湯川，河岡: アバウト推論 -「常識的な推論」を目指して-，AI学会人工知能基礎論研究会，Vol. SIG-FAI-9401-1，pp.1-8 (1994)
2. 笠原，松澤，石川: 国語辞書を利用した日常語の類似性判別，情処論文誌，Vol.38, No.7 (1997)
3. D.B.Lenat, R.V.Guha, K.Pittman, D.Pratt and M.Shepherd: Cyc: Toward Programs with Common Sense, Communications of the ACM, Vol.33, No.8 (1990)
4. 池原，宮崎，白井，横尾，中岩，小倉，大山，林: 日本語語彙大系，岩波書店 (1997)
5. J. Electronic Dictionary Research Institute Ltd. Edr electronic dictionary technical guide (1995)