因果知識と概念ファジイ集合を利用した類推手法の提案

Analogy System using Cause-and-effect And Conceptual Fuzzy Sets

茂呂 佳令[*1] Yoshinori Moro　金 唯錡[*2] Yuki Kin　鈴木 瑠璃[*2] Ruri Suzuki　高木 友博[*1] Tomohiro Takagi

[*1] 明治大学理工学研究科基礎理工学専攻
Computer Science Course, Graduate School of Science and Technology, Meiji University

[*2] 明治大学理工学部情報科学科
Department of Computer Science, Meiji University

Abstract: We propose an analogical inference system which works with causal knowledge base. The system has two main features. First, similar causality is found paying attention to verbal parts. Second, an inference result is generated as a conceptual fuzzy set. Validity of the proposed system is shown through experiments.

1. はじめに

現在計算機可読で大量に蓄えられた知識や経験があり，それらを活用したいという要求がある．知識の中でも「何をした結果，どうなったか」という因果関係の知識の抽出に関する研究が多く行われている[乾 05]．しかし，抽出された因果関係の知識をどのように活用していくかという研究はまだ少ない．因果関係の知識の活用方法として因果関係の知識が抽出された分野以外へと応用をさせるもの，すなわち類推がある．

類推とは人間の行う高度な推論である．通常計算機上で言葉を扱う際に，言葉の意味はほとんど考慮されない．しかし，類推では言葉の意味どころか暗黙的な背景なども考慮する必要がある．そのため，我々はその問題に対処するために概念ファジイ集合[Sekiya 06]と呼ばれる技術を応用した．

本論文では，文書データから抽出された因果関係の知識を，未知の問題に対しても応用が出来る類推を行うシステムを提案する．主な特徴は(1)因果関係にある二つの事象の述語部分に注目して，類似因果の検索を行う，(2)検索した類似因果を概念ファジイ集合を用いて写像を行う．実験により，(1)と(2)が有効かどうかを検証し，その結果から総合的に提案システムの有効性を検証していく．

2. 概念ファジイ集合

2.1 概念ファジイ集合とは

概念ファジイ集合とは，ある語を別の語群の活性値分布で表現したもので，文脈によって変化する語義を表現することを目的として提案された[Takagi 95]．同じ語でも文脈（観点）が変化すれば，概念ファジイ集合もまったく別のものが生成されるという特徴がある．

“peace”と“tokyo”の 2 通りの観点から生成された“hiroshima”の意味を表す概念ファジイ集合を表 1 に示す．表 1 の左側の“peace”という観点からの“hiroshima”では“nagasaki”や“iraq”などを含む戦争にかかわる概念ファジイ集合を生成している．一方，右側の“tokyo”という観点からみた “hiroshima”では“hanshin”や“yakult”などを含む野球にかかわる概念ファジイ集合を生成している．

表 1　観点の異なる概念ファジイ集合の生成

"peace"という観点からの "hiroshima"	
word	value
hiroshima	4390.01
nagasaki	542.96
iraq	191.14
yugoslavia	153.22
the	131.05
dresden	112.11
cambodia	109.05
beijing	108.51
pan	107.67
banghdad	106.14
1945	106.13
tel	102.97
serbia	102.84
tripoli	91.99
japan	90.11

"tokyo"という観点からの "hiroshima"	
word	value
hiroshima	14345.47
hanshin	1059.39
nagasaki	669.39
yakult	662.99
yomiuri	386.89
yokohama	369.13
nagoya	353.97
japan	348.39
front-running	343.67
osaka	251.72
chuunichi	251.63
reshuffled	181.08
convenes	154.76
sendai	147.12
cabinet	130.49

2.2 概念ファジイ集合の写像への応用

類推では，「A→B である．では，A’ →X？」という問題を解く．

（世界1）A → B
　　↓（写像）
（世界2）A’ →X？

X を求めるには A→B の B 部分を A’に合うように写像する必要がある．概念ファジイ集合の文脈によって変化する語義の表現を写像に応用することで，因果関係の知識を入力に合うよう

連絡先：茂呂佳令，明治大学 理工学研究科 基礎理工学専攻，
214-0034 川崎市 多摩区 東三田 1-1-1，
Tel: 044-934-7483，Fax: 044-934-7912，moro@cs.meiji.ac.jp

に写像を行う. 同じ因果関係の知識でも, 入力 A'が変化すれば, 異なる写像が行われる. 同じ語でも観点を変化させることで異なる概念ファジイ集合が生成される点に写像と類似性がある. 概念ファジイ集合を生成する際に, 語の周辺語に注目をしている. 周辺語には類推で写像を行うのに必要な文脈や状況を表す語が多くある. 本研究では, B の観点 A'における概念ファジィ集合を求めることで置き換え写像を実現する.

3. 因果知識データベース

3.1 識別子による因果知識の抽出

因果知識データベースを構築するために, まず因果知識を抽出する. 因果関係にある事象対は明示的な語を間に挟んで文書に出現しやすい. 明示的な語とは「ので」や「ため」など因果関係を示唆する語で, 以下識別子と呼ぶ. 識別子を使い文書から「A 識別子 B」となっている文を抽出する.

図 1 のように係り受けのため, 識別子の直前, 直後の事象対が因果関係にないことが多い. 例えば, 係り受けを無視した因果知識の抽出では「炉の一部を停止するため, 他工場に運び込む」という間違った因果知識を抽出することになる. そのため係り受け解析器 Cabocha を利用し, 識別子に係る事象と, 識別子から係る事象を抽出し, 因果知識とした. そうすることで「炉の一部を停止するため, 他工場に運び込むゴミ量が増えている」という正しい因果知識を抽出することができる.

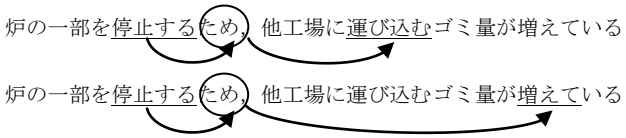

図1 係り受けを考慮した因果知識の抽出

3.2 人手による因果知識データベースの構築

文書データに対して識別子による因果知識の抽出を行った. 文書データは 1990 年の読売新聞の記事 1 年分を使用し, 経済関連の記事(記事分類コード:Y01)のみを使用している. これは景気動向の記事に因果関係の記事が多いとされているためである[坂地 08].

識別子により自動抽出された因果事象対は, 全部で 251 個あった. それに対して 5 人の手作業による正解判定を行い, 3 人以上が因果と判断したものを正解とした. その結果, 因果関事象対候補 251 個中 86 個が正解と判断され, それらを因果知識データベースとした.

4. 学習方法

4.1 日本語の概念ファジイ集合の学習方法

英語では語順が文法の主要メカニズムであり語の区切りが明確なため, 概念ファジイ集合の学習に関矢ら[Sekiya 06]は周辺語として, 語の直前の 4 単語を覚えさせている.

しかし日本語では語の区切りが不明確であるため, 区切り方で大きく結果が変わってしまう. 我々は日本語の区切り方の主流な方法を 5 つ集め, 日本語の概念ファジイ集合の学習に相応しいものを検討した.

(1)形態素 (助詞も含めて学習する)
(2)文節(助詞も含めて学習する)
(3)形態素(助詞を除いて学習する)
(4)文節(助詞を除いて学習する)
(5)N-gram

(3)や(4)のように, 助詞を除いて学習を行うと, 日本語の細かいニュアンスが伝わらないと判断したため, 除外した. また, N-gram は情報検索のように語のマッチングをとる際には有効な手法だが, 言葉の意味を捉えて区切る方法には不向きだと考え除外した.

(1)と(2)で概念ファジイ集合生成の予備実験を行った. その結果, (1)の形態素で区切った場合は, こそあど言葉, 助詞, 漢数字が概念ファジイ集合の生成結果の大半を占めてしまい, 正しい生成結果が得られなかった. (2)の文節で区切った場合は, 直前の文が長くなり, 一致条件がシビアになった. そのため生成結果がない, もしくは極めて少ないという結果になった.

以上の結果より, 一般的な日本語の区切り方である, 形態素区切りや文節区切りのどちらでも, 概念ファジイ集合が生成されないことが判明した. そのため, 我々は形態素区切りと文節区切りを合わせた最小文節区切りを提唱する.

4.2 最小文節区切り

最小文節区切りは, 形態素区切りと文節区切りの中間に近い区切り方である. 最小文節区切り方の定義は「基本的には形態素で区切る. ただし, 区切った形態素が助詞の場合は前の形態素に連結する」である. 以下具体例を示す.

文 : 郵政民営化を見直す
形態素 : 郵政／民営／化／見直す／
文節 : 郵政民営化を／見直す／
最小文節 : 郵政／民営／化を／見直す／

例で示されるように, 最小文節で区切ることで, 概念ファジイ集合の学習時の形態素区切りの際の助詞が単独で出現する問題と, 文節区切りの熟語が長くなりすぎる問題を解決することができる.

また, 最小文節で区切ることで, 概念ファジイ集合の Modified Revised Confabulation Model[Sekiya 06]で「最小文節＋最小文節」のまとまりが文節になる場合が多く, 次に出現する語を予測する際に, 人に近い感覚で予測することができる.

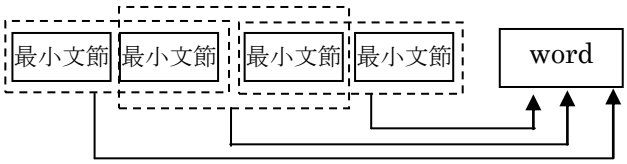

図 2 日本語における概念ファジイ集合の生成

5. 因果関係類推システム

5.1 因果関係類推システムの概要

図3に因果関係類推システムの概要を示す. 入力原因と類似する原因を因果知識データベース(因果知識 DB)から検索する. 検索した類似原因と対となる結果(起こりうる結果)を概念ファジィ集合を用いて生成する. このとき, 因果知識データベース内の1つの原因に対して, 起こりうる結果は通常複数存在する.

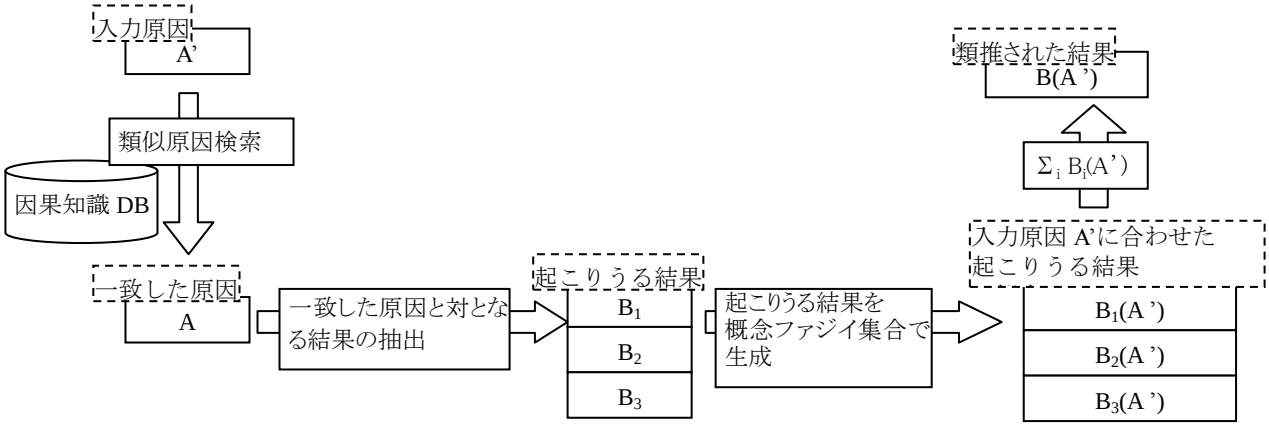

図3 因果関係類推システムの概要図

5.2 類似原因検索

因果関係が似ているかどうかの判断は困難である. 例えば,「子供が転んだ」と「ロボットが倒れた」は人が見れば似た事象であると判別がつくが, 計算機にその判断をさせることは非常に困難である.

乾ら[4]は「共通の述語を持つ事象ペアは因果関係の有無に関しても, 同様の傾向にある」と仮定している. 同様に我々も「共通の述語を持つ事象ペアは, 因果関係の内容に関しても同様である」と仮定している. 例えば「子供が転んだ」と「自転車で転んだ」は異なる事象であるが述語が「転んだ」と共通であるため, 似た事象としている. この仮定により類似原因検索がシンプルになる. 因果関係データベースから類似原因を検索するときは事象の述語が一致してるかどうかのみに注目をすればよい. この仮定を満たさない例外に関しては, 後述の概念ファジイ集合を利用することで問題を解決する.

5.3 因果関係類推システムにおける概念ファジイ集合の役割

5. 2 節に示したように類似原因検索を行いやすくするために, 因果の事象対の述語部分を利用している. 概念ファジイ集合は「子供が転んだ」のような文のような形では生成することができない. ここでも同様に「転んだ」のような述語部分を利用する.

本システムの概念ファジイ集合の役割は 2 つある. 1 つ目は似た結果を持つ概念ファジイ集合をマージすることで, 起こりやすい結果を強める効果を持つことである. 例えば,「転ぶ」と「倒れる」は同じような概念ファジイ集合を生成する. 複数の起こりうる結果の概念ファジイ集合をマージすることで, 共通して起こりやすい結果を強調することできる. 2 つ目はノイズフィルターである. 述語の一致のみでは, 5. 2 節での仮定を満たさない例外(ノイズ)が存在する. その問題は概念ファジイ集合を生成する際に, 観点を利用することで解決する. 観点となる入力された原因と起こりうる結果の関連度が低くなるため, 概念ファジイ集合が生成されにくい. そのため, 観点に関連のある起こりうる結果のみから概念ファジイ集合が生成される.

5.4 概念ファジイ集合からの文の合成

概念ファジイ集合は語の集合であるため, 人にとって読みづらくなっている. そのため, それらの語を組み合わせることで文の合成を行う. 概念ファジイ集合を述語集合と述語以外の集合(主に名詞や副詞, 形容詞)の二つに分け, 相互情報を用いて関連度の計算を行い, 関連度の高い順に文を合成した.

$$M = \log\left(1+\frac{P(x,y)}{P(x)\times P(y)}\right) \quad \ldots\ldots(1)$$

$P(x)$: 述語が文書データに出現する確率
$P(y)$: 述語以外の語が文書データに出現する確率
$P(x,y)$: 述語と述語以外の語が同時に文書データに出現する確率

6. 検証

6.1 検証内容

本来なら, 類似原因検索が適切に行われているか, 概念ファジイ集合による写像を適切に行われているかどうかの 2 点を検証する必要がある. しかし, 本システムは写像を行う際の概念ファジイ集合の生成が類似原因検索の内容に干渉しているため, その検証を行うことができない.

そのため, 類推システムに手動で作成した原因を入力として与え, 正しい結果を返すことができるかどうかのみを検証していく. 原因は次の点に注目して作成した. 1 つ目は原因と対となる結果が 1 つになるもの. これは対となる結果が 1 つだと概念ファジイ集合を生成した際に複数集合のマージが必要なく, 概念ファジイ集合の生成の評価を行いやすいためである. 2 つ目は原因と対となる結果が複数になるもの. これは複数の対となる結果の概念ファジイ集合をマージした結果の評価を行うためである.

6.2 検証環境

学習用の文書データには 1990 年の読売新聞を使用した. 学習には 1 年分の記事を全て使用している. 因果知識データベースには 3. 2 節で述べたように経済関連の記事(記事分類コード: Y01)を使用し, その後人手で正例のみを抽出した.

6.3 検証

対となる結果が 1 つになる原因として「輸入品が値上がりした」を類推システムに入力した. 因果関係データベースで述語部分の「値上がりした」と対となる結果は「上昇する」「なる」である. このとき高頻出語から概念ファジイ集合を生成した際の精度が低くなる傾向にあるため,「なる」は対となる結果から除去した.

その出力結果が表 2 である. 述語集合では「上昇」や「高騰」の語が多く出力された. 述語以外の集合では「与野党」や「法案」などの政治に関わる語が多く出力された. それらを合成結果では上昇関連の語や政治関連の語が多く出力された. しかし,「大幅に上昇した」のように主語が存在しないものや「法案のめぐる」のようにどのような法案なのか情報不足のものが多く出力された.

表 2　類推システムの出力結果 上位 15 件 （対となる結果が 1 つのもの）

入力原因
輸入品が値上がりした

入力原因の述語部分と対となっている結果
上昇する
なる

述語以外の集合	
word	value
与野党	0.03195
税	0.03173
野党	0.03062
税制	0.02914
消費	0.02870
上昇	0.02711
大幅に	0.02662
大幅	0.02510
見直し	0.02487
代	0.02482
政治	0.02345
法案の	0.02326
政府	0.02241
廃止	0.02240
円	0.02183

述語集合	
word	value
上昇した	0.04399
上昇して	0.03693
上昇する	0.03189
上昇し	0.02620
という	0.02582
つける	0.02075
高騰して	0.02039
急騰し	0.02019
による	0.01840
上昇すれば	0.01767
高騰し	0.01750
下回った	0.01578
めぐる	0.01556
反落した	0.01500
下落して	0.01485

合成結果	
word	value
大幅に上昇する	3.77961
上昇上昇する	3.75312
大幅下落する	3.57325
大幅上昇する	3.45183
大幅に下回る	3.44412
与野党めぐる	3.34533
大幅に高騰する	3.30592
上昇高騰する	3.27857
大幅反落する	3.08097
上昇下落する	2.96191
大幅に下落する	2.67233
税制めぐる	2.52941
法案のめぐる	2.45784
上昇下回る	2.31919
見直し上昇する	2.31061

人が見れば，「大幅に上昇した」は「物価が大幅に上昇した」というように予測できるものが多い．対となる結果が 1 つの場合では，「物価が上昇した」のような明確な結果を出力することができず，課題を多く残す結果となった．

対となる結果が複数になる原因として「輸入が順調な伸びを示す」を類推システムに入力した．因果関係データベースで述語部分の「示す」と対となる結果は「求める」「注目する」「迫る」「なる」である．このとき高頻出語の「なる」を除いた 3 つの対となる結果から概念ファジイ集合を生成マージした．その結果は，対となる結果が 1 つのときよりも悪かった．

本来，複数の概念ファジイ集合をマージすることの目的は多くの概念ファジイ集合に出現する結果を強調するためである．しかし，実際に複数の概念ファジイ集合をマージすると，それぞれの概念ファジイ集合に出現するノイズを上位の結果に出してしまう．その結果，複数の概念ファジイ集合をマージすることで，より精度が下がる結果となった．

6.4 まとめ

類似原因検索や概念ファジイ集合の生成では事象の述語部分のみに注目している．因果関係において「何が起きたか」を知ることは可能であるが，類推をする上では不十分であった．また述語部分に「する」，「いる」，「なる」などの高頻出語が多く，それらの語から概念ファジイ集合を生成することが難しく，類推に使用することができなかった．概念ファジイ集合を写像に応用することは可能であるが，概念ファジイ集合の学習方法の改良や学習量を増やすことが必須である．現段階で日本語の概念ファジイ集合の生成の精度が低いため，複数の概念ファジイ集合をマージすることで精度の低下を招いてしまっている．今後，概念ファジイ集合の生成の精度が向上させることで，より良い結果を出すことは可能であると予想する．

参考文献

[乾 03] 乾孝司，乾健太郎，松本裕治: テキストから獲得可能な因果関係知識の類別およびその自動獲得の試み -接続助詞「ため」を含む文を中心に-，言語処理学会第 9 回年次大会，pp.707--710，2003.

[乾 05] 乾孝司，乾健太郎，松本裕治: 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，情報処理学会論文誌 Vol.45，No.3，pp.919--933，2004.

[乾 05] 乾孝司，奥村学: 文書内に現れる因果関係の出現特性調査，計量国語学，Vol.25，No.3，pp.123--144，2005.

[乾 06] 乾孝司，高村大也，奥村学: 因果関係知識獲得のための隠れ変数モデル，言語処理学会第 12 回年次大会，pp.959--962，2006.

[坂地 08]坂地泰紀，竹内康介，関根聡，増山繁: 構文パターンを用いた因果関係の抽出，言語処理学会第 14 回年次大会，pp.1144-1147，2008.

[Sekiya 06] H. Sekiya，T. Kondo，M. Hashimoto，and T. Takagi: Dynamic Sense Representation Using Conceptual Fuzzy Sets，Journal of Advanced Computational Intelligence and Intelligent Informatics，Vol. 10，No. 6，859-867，2006.

[Sekiya 06] H. Sekiya，T. Kondo，M. Hashimoto，and T. Takagi: Conceptual Fussy Sets Generation Depending，Proceedings of North American Fuzzy Information Processing Society，2006.

[Takagi 95] T. Takagi，A. Imura，H. Ushida，and T. Yamaguchi: Conceptual fuzzy sets as a meaning representation and their inductive construction. International，Journal of Intelligent Systems，Vol. 10，pp. 929.945，1995.