パターンと機械学習を用いた大規模テキストからの変遷情報の抽出と分類

堀さな子 [*1]　村田真樹 [*1]　徳久雅人 [*1]　馬青 [*2]
[*1] 鳥取大学 工学部 知能情報工学科
[*2] 龍谷大学 理工学部 数理情報学科
{s072048,murata,tokuhisa}@ike.tottori-u.ac.jp
qma@math.ryukoku.ac.jp

1　はじめに

物事の変遷を知ることはその物事の知識を会得する時に重要なことである．この変遷を知るためには一般的に Web や検索エンジン，または書籍を使用して情報を得る方法があげられるが，これらの方法では人手では網羅的に収集するのが困難であり，かつ多大な労力を要する．変遷を知ることを自動で簡単に行えれば便利であり，また手間が省ける．

我々の先行研究 [1] ではその手始めとして，論文のタイトル，著者名のデータを使用し，どの分野がどの分野から発生したか，どの研究者がどの研究者を教示していたかの関係 (いわゆる先生と弟子のような関係) を自動で推定した．次に，Fan ら [2] は法則の変遷情報を Wikipedia から抽出した．法則ページ (法則を記載したページ) に記載されている年号より各法則の発見年を予測し，ある法則 A のページに他の法則 B が記載されている場合に法則 A と法則 B が変遷の関係にある可能性が高いとするヒューリスティックルールに基づき，法則 A と法則 B の対をそれぞれの法則の発見年とともに変遷情報として抽出した．

これらの研究は，問題点として学術分野間，師弟間，法則間という限定された変遷の種類についての抽出であったことが挙げられる．それに対する課題として以下がある．

- より多くの種類の変遷を自動で取得すること．
- 取得できた変遷の種類を明らかにすること．
- 多くの種類の変遷の取得を自動でかつ高精度で行うこと．また変遷の種類も自動で推定すること．

本稿では，これらの問題を解決するために以下の研究を行う．

1. 大量の文から人手で作成したパターンを利用し，変遷情報を自動で抽出する．(3 節)
2. 1 で抽出した変遷情報を人手で分類し，分析する．(4 節)
3. 2 で人手で分類したデータを教師データとして教師あり機械学習を利用してより高性能に変遷情報の抽出を行う．さらに変遷情報の自動分類も行う．(5 節)

本研究の主張点を以下に整理する．

- 大規模テキストから変遷情報を取り出すという特色のある研究対象を扱った．
- 変遷情報を，文での変遷情報の含み方，変遷情報の分類，変遷情報における概念の変化の仕方の 3 つの観点で分類した．これは変遷情報を扱う際の理論的基礎として今後役立つと考える．
- パターンで変遷情報を含む可能性のある個所を抜き出し，そこから機械学習でより高性能に変遷情報を抜き出す手法を提案した．この手法はパターンを用いるだけの手法よりも性能が高いことを確認した．本研究の実験において提案手法は 0.9 という高い F 値を得た．

2　変遷情報

本研究において変遷とは，二つの事物 X，Y において，X が時の流れとともに移り変わり Y となった，または，X が影響を及ぼして Y になったという変化のことを変遷とする．そして，X → Y と表記し，X，Y のペアを変遷情報と呼ぶ．

3　パターンに基づく変遷情報の抽出

3.1　提案手法

大量の文から人手で作成したパターンを利用して，変遷情報を自動で抽出する．この手法はパターンを用いて行っているため，以後「パターンベース法」と呼ぶ．本研究では，パターンに基づく自動抽出に，ALAGIN の「意味的関係抽出サービス」[3] を利用する．これは，Stijn らの研究 [4] を元にしたサービスである．

このサービスでは，「原因-結果」「トラブル-予防策」「食材-効能」などの種々の意味的関係を持つ単語対を作成することができる．統計的な手法を用いた半自動処理により，約 6 億の Web 文書から効率的に大量の単語対を抽出することができる．このサービスでは「X から派生する Y」などのパターン (シードパターン) を入力すると，シードパターンと同様な意味関係を持つ類似したパターンを自動で作成し，シードパターンと自動で作成した類似パターンに合致した X，Y を Web 文書から自動的に抽出する．

本研究での事前の実験では，シードパターンから作成した類似パターンを利用すると，抽出されたものの性能は低かった．このため，本研究では自動で作成される類似パターンは利用せず，シードパターンのみを利用する．

変遷情報の取得に役立つパターンを人手で作成する．そのパターンを上記サービスで利用することで，パターンに合致する文と X，Y にあたる名詞を変遷情報として取り出せる．以下に利用するパターンの例を載せる．

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

シードパターンの例

<X から生まれた Y>
<X を元にした Y>
<X の元である Y>
...

3.2 実験と結果

シードパターンは 34 個入力した．以下に抽出できた文の例を載せる．

抽出できた文の例

しかしこのヒドロキシルラジカルの元となる過酸化水素を除去する酵素も人間は持っています。

この実験で抽出できた文は 3,115 文である．しかし，3,000 文程度では少ないと思われる．これはシードパターンが少なかったこと，シードパターンの質が悪かったことが原因として挙げられる．解決方法としてはシードパターンの見直しが考えられる．また，サービスの類似パターンを利用することも考えられる．

4 人手に基づく変遷情報の分類と分析

4.1 提案手法

4.1.1 変遷情報の含み方に基づく分類

3 節でパターンベース法を用いて抽出した文を以下の typeA〜F に分類する．この分類は変遷か否かの判定を明確にする目的を持つ．なお，本稿においては，「X から派生する Y」などのパターンベース法でのシードパターンで X と Y に当てはまる名詞により，X と Y が変遷関係にあるかを判断する．

表 1 に例を示す．下線部がシードパターンに当てはまっているものであり，二重線部の単語が X と Y にあたる単語である．

type-A X，Y が明らかに変遷情報であり，知見の得られる事例．ただし，X，Y 自体が変遷関係にない場合であっても，X，Y に対して修飾関係 (連接した修飾関係) にある語が変遷関係にある場合も type-A とする．

type-B X,Y のどちらか一方が一般的に広い意味を持つ名詞であるが，文の構造からその名詞の具体的内容を示す表現がその文の他の個所から抽出できる事例．

type-C X,Y のどちらか一方が一般的に広い意味を持つ名詞であるが，X，Y の名詞から変遷として知見の得られる事例

type-D X，Y のどちらか一方，もしくは両方が一般的に広い意味を持つ名詞であり，変遷として知見の得られない事例

type-E 単に場所を指定している事例

type-F 単に状態を表している事例

type-A〜C が変遷情報の取得に役立ち，それ以外は変遷情報の取得に役立たない．

4.1.2 変遷の種類に関する分類

抽出した変遷を以下の分類 1〜8 に分ける．この分類は，大規模なテキストから抽出した広範な内容を含む変遷の文に，どのような種類の変遷が含まれているかを明らかにする目的で行う．表 2 に例を示す．分類 1〜8 の定義は事前に Wikipedia からパターンで取り出した変遷情報を含む文を人手で分析して作成した．

分類 1 方法・装置・道具等により変遷がなされた事例

分類 2 文化，芸術などの分野からの変遷を示す事例

分類 3 発想，思想，考えなどから変遷したものを示す事例

分類 4 原因と結果の関係を示す変遷の事例

分類 5 名称の変遷を示す事例

分類 6 成分・原料などの構成要素とそれから構成されるものの関係を示す変遷の事例

分類 7 ある生物とその元の生物や成分などとの関係を示す事例

分類 8 その他の事例

表 1: 変遷情報の含み方に基づく分類ごとの文の例

分類	文の例
type-A	発生過程を再現するように、ＥＳ細胞を神経 (Y)の元になる幹細胞 (X) などに分化させ、さらに条件を変えて培養することで、前脳型アセチルコリン作動性神経細胞など様々な神経細胞を作り分けることに成功した。 (解説：「幹細胞」は「神経」の元の物質であるため，変遷とみなせる
type-B	精米の目的は、お米の表面近くに分布する、タンパク質や粗脂肪などのお酒の雑味 (Y)の元となる成分 (X) を取り除くことにあります。 (解説：「成分」は一般的に広い意味を持つ名詞であるが「タンパク質や粗脂肪などの」で修飾されている)
type-C	臭いや にきび (Y)の元となる原因菌 (X) の殺菌効果に優れたボディソープです。 (解説：「原因菌」は一般的に広い意味を持つ名詞であるが X と Y を 2 つ見て知見が得られる)
type-D	ＪＡＡＡ３０年の歴史は、まさに「自動化に絡む 企業活動 (X)から派生する関係性 (Y) を楽しむ充足感」を動機として運営されてきた (解説：「関係性」は一般的に広い意味を持つ名詞であり，知見が得られない．)
type-E	また同社はブルゴーニュ全域でも最大の土地所有者のひとつに数えられ、１００ｈａの自社畑 (X)から生まれるワイン (Y) は、同社の生産量の８５％を占める。 (解説：「自社畑」は場所を示している．)
type-F	同じような境遇 (X)で生まれた組織の先輩 (Y) には『００７美しき獲物たち』の悪役、マックス・ゾリンがいます。 (解説：単に「組織の先輩」の状態を表している)

4.1.3 変化の仕方に基づく分類

抽出された変遷情報の X,Y がどのように変化をしているかに基づき，変遷情報を以下の 6 種類に分類する (図 1)．表 3 に例を示す．

change X が Y に変化した事例．

part・x X の一部が Y に変化した事例．

part・y X が Y の一部に変化した事例．

part-part X の一部が Y の一部に変化した事例．

effect X が何らかの影響や作用をし何かが Y になっ

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表 2: 変遷の種類に関する分類ごとの文の例

分類	文の例
分類 1	遺伝子操作 (X)で生まれた新人類 (Y)「コーディネイター」と ロボット工学 (X)で生まれた人造人間 (Y)「ヒューマノイド」は、戦えば、どっちが強い？
分類 2	着物文化 (X)から生まれた布づくり (Y) の伝統を、豊かな発想でジャパニーズテキスタイルとして進化させながら、世界に発信している宮本英治。
分類 3	ハンドクリーム発想 (X)から生まれた消毒ジェル (Y)。
分類 4	昼にビールを・・とお考えの方も多いと思いますが、アルコールは 疲労 (Y)の元となる尿酸 (X) の排出を妨げ、尿酸値を高めるため、お勧めは出来ません。
分類 5	日進は四字成語「日進月歩」からとられたもので、最初に「日進小学校」に付けられ、後から日進村・村の名 (合併して字の名) となり、「日進駅」(Y)へ派生した駅名 (X) である.
分類 6	焼酎 (Y)の原料であるサツマイモ (X) は全て地元契約農家に栽培を委託しております。
分類 7	彼らは群れて住む邪悪な生物であり、卵 (X)から生まれる哺乳類 (Y) で、頭には角が生え、ウロコがあり、その姿はトカゲを想わせる醜い二足動物とされている。
分類 8	それに対して、ボランティア活動 (X)で生まれた関東の某市長 (Y) さんは、名古屋での講演の後みんなでわいわいガヤガヤとお喋りした後、…

表 3: 変化の仕方に基づく分類ごとの文の例

分類	文の例
change	赤ワインの 渋みや旨み (Y)の素であるポリフェノール (X) の成分の一つで、空気に触れるとエステルに変化し、口当たりが良くなるといわれている。
part・x	コラーゲン (Y)の材料となるウロコ (X) を時間内にいかに多く集められるかを競う、シンプルながら手に汗握るアクションゲーム！
part・y	エビやカニなど甲殻類や昆虫の外皮 (Y)の成分であるキチン (X) もセルロースと並ぶ生体構造ポリマーです。
part-part	酒税法では、ラム酒 (Y)の原料であるサトウキビ (X) の搾り汁や糖蜜を使った焼酎造りは禁止されているが、米国統治から日本に復帰した昭和２８年、米麹の使用を条件に奄美群島に限り特例として認められた。
effect	バイオ技術 (X)から生まれたアガリクス菌糸体 (Y) の驚異的力
none	広さ１２ｈａの小規模な 葡萄園 (X)から生まれるワイン (Y) は、軽快でエレガントなタイプのもの。

表 4: 変遷情報の含み方に基づく分類の結果

分類	type-A	type-B	type-C	type-D	type-E	type-F
個数	65/100	6/100	5 /100	17/100	5/100	2/100

た事例.

none 関連性のない事例.

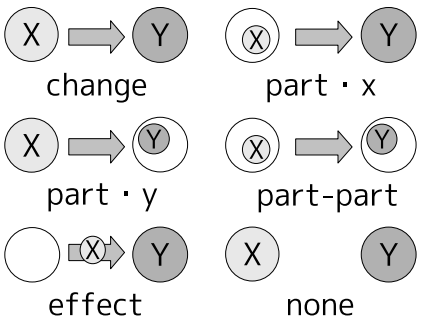

図 1: 変化の仕方に基づく分類

4.1.2 節の変遷の種類に関する分類ごとに、変化の仕方の特徴があると思われる．4.1.2 節の分類とあわせて分析することで、変遷の分類ごとの変化の仕方の特徴を見い出す.

4.2 実験と結果

3 節のパターンベース法により抽出した文からランダムで 100 文取りだし，4.1.1 節の typeA～F に分類した.

結果を 4 に載せる.

100 文中の 6 割以上が typeA に分類され，変遷と分類された．また，typeA～C を変遷とみなせると考えると，文の数は 76 個となった.

また，typeA～C である事例 76 個を，4.1.2 節の分類 1～8 に分類した結果を表 5 に示す.

表 5: 変遷の種類に関する分類の結果

分類	分類 1	分類 2	分類 3	分類 4	分類 5	分類 6	分類 7	分類 8
個数	7/76	0/76	0/76	29/76	0/76	36/76	2/76	2/76

結果としては，分類 6 の文数が一番多く，分類 2，3，5 については分類される文がなかった.

typeA～C である 76 個の事例を，4.1.3 節の変化の仕方で分類した結果を表 6 に示す．なお，分類 2，3，5 については 100 文のうちに存在しなかったため載せていない.

表 6: 変化の仕方の分類結果

変化の仕方	分類 1	分類 4	分類 6	分類 7	分類 8	合計
change	0/7	19/29	5/36	2/2	0/2	26/76
part・x	0/7	0/29	3/36	0/2	1/2	4/76
part・y	0/7	0/29	26/36	0/2	1/2	29/76
part-part	0/7	0/29	2/36	0/2	0/2	0/76
effect	7/7	10/29	0/36	0/2	0/2	17/76
none	0/7	0/29	0/36	0/2	0/2	0/76

分類 1 の場合，effet の変化の仕方を示すことがわかった．分類 1 ～8 ごとに分類の変化の仕方に特徴があり，それを確認することで人手に基づく分類 1～8 および変

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

化の仕方の分類付与の誤りの検出にも役立った．

5 機械学習に基づく変遷情報の抽出と分類

5.1 機械学習に基づく変遷情報の抽出

5.1.1 提案手法

3 節のパターンベース法に加えて機械学習を用いることでより高性能に変遷情報を抽出する．パターンベース法で取得したものが真に変遷情報であるかを教師あり機械学習で判定し，機械学習が変遷情報と判断したもののみを変遷情報として取り出す．機械学習には SVM，最大エントロピー法 (ME) を使用する．機械学習の素性として，表 7 に示すものを用いる．

表 7: 機械学習の素性

素性番号	素性の説明
1	合致したパターン
2	ALAGIN のサービスの出力する名詞 X のクラス番号
3	ALAGIN のサービスの出力する名詞 Y のクラス番号
4	X の名詞
5	Y の名詞
6	名詞 X の直前の単語
7	名詞 X の直後の単語
8	名詞 Y の直前の単語
9	名詞 Y の直後の単語
10	同一文中の名詞，動詞，連体詞，副詞，形容詞，接続詞，感動詞

5.1.2 実験と結果

パターンベース法を使用して抽出した 100 文のデータを用いて，変遷が含まれているかの判定を機械学習で行う．実験は 10 分割のクロスバリデーションで行っている．データは 4.1.1 節の方法で分類したものを用いる．typeA〜C を変遷であるもの，typeD〜F を変遷ではないものとして判定を行ったものを使用した．

表 8 に結果を載せる．ここでの再現率の分母はパターンベース法で抽出した 100 文を利用している．なお，表 8 の「パターンベース法＋ SVM」は 3 節のパターンベース法により変遷の候補を取りだし，SVM により変遷が含まれているかの判定を行った方法，「パターンベース法＋ ME」は 3 節のパターンベース法により変遷の候補を取りだし．ME により変遷が含まれているかの判定を行った方法である．

表 8: 変遷情報の自動抽出の性能

方法	再現率	適合率	F 値	正解率
パターンベース法のみ	1.00	0.76	0.86	0.76
パターンベース法＋ SVM	0.92	0.90	0.91	0.87
パターンベース法＋ ME	0.93	0.86	0.89	0.84

結果としては，SVM，ME を追加して行った方法の両方でパターンベース法のみの方法よりも F 値が上昇した．変遷を抽出する際，機械学習を使用することが効果的であることが確認できた．

5.2 機械学習に基づく変遷情報の分類

5.2.1 提案手法

4.1.2 節の方法で分類したデータを使用し，機械学習を用いて変遷情報の種類の分類を行う．素性は 5.1.1 節と同じものを用いる．

変遷情報の種類の分類が自動で行うことができれば，取得した変遷情報を効果的に分類でき，便利と考える．

5.2.2 実験と結果

10 分割のクロスバリデーションで実験を行った．結果を表 9 に載せる．

なお，分類 2，3，5 については 100 文のうちに存在しなかったため載せていない．

表 9: 変遷の種類に関する自動分類の性能

	SVM				ME			
分類	総数	再現率	適合率	F 値	総数	再現率	適合率	F 値
分類 1	7	0.28	0.40	0.33	7	0.57	0.66	0.61
分類 4	29	0.58	0.68	0.62	29	0.62	0.66	0.64
分類 6	36	0.72	0.56	0.63	36	0.72	0.60	0.65
分類 7	2	0.00	—	—	2	0.00	—	—
分類 8	2	0.00	—	—	2	0.00	—	—

結果としては，正解率は SVM で 0.59，ME で 0.63 であった．F 値は ME の方が良い．この機械学習による分類では，分類された文の総数が多いものでは F 値が 6 割以上あり良いが，少ないものでは性能が低いという結果となった．これは文の数を増やせば性能は向上すると思われる．

6 おわりに

本稿では，大量の文からパターンに合致するものを取得することで，幅広い分野の変遷情報を取得した．この方法で 0.86 という高い F 値で変遷情報を抽出できた．また，抽出した変遷情報を分類し，どのような種類の変遷があるのか知見を得ることができた．例えば，今回の研究では成分・原料などの構成要素とそれから構成されるものの関係を示す変遷の事例が，一番多い変遷の分類であることがわかった．更に，機械学習 (SVM,ME) を使用し，より性能高く変遷を抽出することを行った．実験の結果，SVM では F 値 0.91，ME では F 値 0.89 で変遷を抽出できた．本研究により，パターンに基づく方法と機械学習を組み合わせることで，より性能高く変遷を取り出せることがわかった．機械学習による変遷情報の分類では，学習データの事例数の多い分類では F 値が 6 割以上であった．

参考文献

[1] Sanako Hori, Masaki Murata, Masato Tokuhisa, Qing Ma: Automatic extraction of historical transition in researchers and research topics. NLPKE 2011: 296-299,2011.

[2] Liangliang Fan, Masaki Murata, Masato Tokuhisa and Qing Ma: Extraction of Historical Transition in Legal and Scientific Laws from Wikipedia. NLPKE 2012 : pp.144-155, 2012.

[3] 高度言語情報融合フォーラム: “ALAGIN 意味的関係抽出サービス”，https://alaginrc.nict.go.jp/

[4] Stijn De Saeger, 鳥澤 健太郎, 風間 淳一, 黒田 航, 村田 真樹: 単語の意味クラスを用いたパターン学習による大規模な意味的関係獲得, 言語処理学会第１６回年次大会 D4-2, pp.932-935, 2010.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.