言語パタンと統計的共起尺度による属性関係抽出*

高橋 哲朗　乾 健太郎　松本 裕治
奈良先端科学技術大学院大学情報科学研究科
{tetsu-ta, inui, matsu}@is.aist-nara.ac.jp

1 はじめに

質問応答タスクにおいて WWW などの大規模文書集合を知識源として用いる場合，情報の冗長性 (redundancy) が有効に働くことが報告されている [7, 2]．冗長性に依る手法では，入力された質問文に対する正しい回答は文書集合中に繰り返し出現しているという仮定に基づき，その出現頻度を用いて回答が選択される．このようなアプローチは WWW などの大規模な文書集合を知識源としている場合には有効であるが，限られた情報源に対する適用は難しい．また大規模文書集合を知識源としている場合でも，出現回数の少ない正解の発見が困難であるという問題がある．

冗長性に頼らない手法としては，あらかじめ文書集合からオフラインで情報を抽出し質問応答のためのデータベースを作っておき，そこから回答を検索するアプローチが考えられる．Lin ら [8] はユーザの質問の種類の数と問われる回数は Zipf's Law に従うという調査結果を示している．すなわち稀にしか聞かれない質問の種類は多いが頻繁に聞かれる質問ほどその種類は少ないので，その種類の質問に対してはオフラインで構築したデータベースの利用が有用であると言える．筆者らが QAC2 [4] の質問集合について質問の種類を調査した結果，全質問中の 24% が属性を直接尋ねる質問であり，間接的に属性を尋ねる質問を含めるとその割合は質問全体の 40% であった [11]．

そこで本研究は質問応答のための静的な知識獲得を目的とし，文書集合から（対象物，属性名，属性値）の三つ組を網羅的に抽出するタスクに取り組む．たとえば「富士山は標高 3776m の山だ」というテキストからは，(**対象物：**富士山，**属性名：**標高，**属性値：**3776m) の三つ組を抽出する．

2 関連研究

属性関係を抽出するタスクはテキストから特定の情報を抽出する問題の一つと見なせるので，これまで Message Understanding Conference （MUC）を中心に行なわれてきた情報抽出（IE）との関連が強い．本タスクは名詞句とそれらの間の関係を抽出するという点において MUC と共通しているが，(a) シナリオを限定しない，(b) 属性関係にある三つ組のみを抽出する，という点において異なっている．

MUC における IE ではテキストからあらかじめ指定されたシナリオに関する情報を抽出することが目的であった．シナリオが限定されている場合は，それに特化した抽出パタンや辞書の作成または獲得が可能である [1, 9]．しかし今回提案したタスクではシナリオやドメインを限定しないため，このようなアプローチをとることが難しい．また情報抽出に機械学習を適用するアプローチも提案されておりその有効性が示されているが [12, 3]，本タスクに同様の手法を適用するためには抽出したい関係ごとにトレーニングデータを用意し分類器を作成する必要があり，そのコストのために本タスクへの機械学習の適用は難しいと言える．

MUC に代表される情報抽出ではシナリオごとに取るべき情報が与えられていた．本タスクではシナリオを限定せずに，あらゆるドメインを対象とするという点で MUC の問題の一般化となっており問題の範囲を広げているが，抽出対象を属性関係だけに限定するという制限を加えることにより問題の範囲を限定している．

Hasegawa ら [5] はシナリオを限定せずに関係を抽出する手法を提案している．彼らの手法では固有表現 (NE) の対をそれらの間にある文脈によりクラスタリングし，それぞれのクラスタ内で文脈中に共通に出現する語を関係名とすることで関係を抽出している．この手法は NE の対を手がかりとして新しい知識を抽出するアプローチであり，そのため対象物と属性値が頻繁に共起していなければ関係の抽出ができない枠組となっている．それに対し我々は対象物と属性値の共起の統計量は用いず，これらの共起頻度が少ない場合でも属性名を仲介として抽出できる枠組を提案する．

3 提案手法

ドメインを限定しない本研究の課題設定では，特定の属性を特徴付けるようなパタンではなく属性名を横断するようなパタンを用いるべきである．そこで本研究では，

- 抽象化した抽出パタンにより属性関係候補を抽出
- 抽象化した抽出パタンにより生じるノイズを統計量を用いてフィルタリング

というアプローチをとる．

概要を図 1 に示す．手順は以下の通りである．

<1> 特定のドメインに依存しない抽象的なパタンを用いて，コーパスから三つ組の候補を抽出する．

*Automatic Extraction of Attribute Relations Using Patterns and Statistical Cooccurrence.
TAKAHASHI Tetsuro, INUI Kentaro and MATSUMOTO Yuji. Nara Institute of Science and Technology

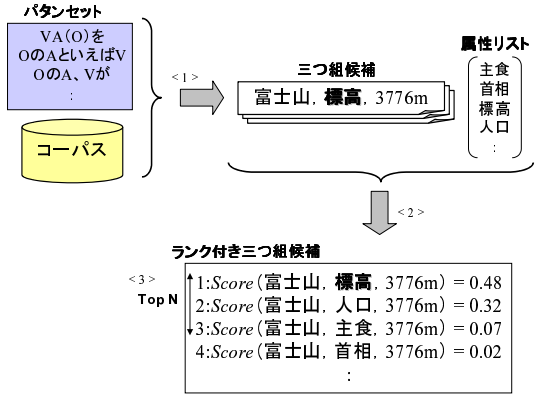

図 1: アプローチの概要

<**2**> パタンで抽出した対象物 (O) と属性値 (V) に対して，あらゆる属性名 (A) において 式 (1) により三つ組のスコア $Score(O, A, V)$ を求める．

<**3**> 推定した三つ組をランキングし，パタンによって求めた三つ組が上位 N 位以内になければ棄却する．

$$Score(O, A, V) = S_{OA}(O, A) \times S_{VA}(V, A) \qquad (1)$$

ここで $S_{OA}(O, A)$ は $O(bject)$ が $A(ttribute)$ を属性としてどれくらい持ちやすいか，$S_{VA}(V, A)$ は $V(alue)$ が $A(ttribute)$ の値としてどれくらい適切かを表す値である．例えば，S_{OA}(富士山, 標高) は「富士山」がどれくらい 「標高」という属性を持ちうるかを表し，S_{VA}(3776m, 標高) は,「3776m」がどれくらい「標高」の値となりうるかを表す．

提案手法では，三つ組の出現に関するスコアを（対象物と属性名）と（属性名と属性値）の 2 つに分けて用いることによりデータスパースネスの問題を回避する．提案手法は，(a) パタンのいずれかにマッチしなければならない，(b) 与えられた文脈中の対象物と属性値から推定される属性名のリストの上位 N 位以内に入っていなければならない，という 2 つのフィルタを用い，これらのフィルタを通過する候補の積集合を求めているという見方もできる．言い換えるとこれら両方の手がかりに支持された三つ組が属性関係として抽出されることになる．

4 評価実験

4.1 タスク設定

本タスクでは対象物を IREX [6] で定義された NE に限定する．また抽出すべき属性の種類は有限だと考えられるので，図 1 中の属性リストもあらかじめ与える．この制約のもとで文書中に現れる属性関係を網羅的に抽出することを目標とする．

属性名のリストを作成するために，まず “NE の x” というパタンをコーパスに適用し x に入る語を抽出した．そしてそれらの語を人手で選別し 1038 種類の属性名からなる属性リストを作成した．

4.2 抽出パタンの作成

抽出パタンを作成するために，まず人手により約 70 個の属性関係の三つ組をシードとして用意した．これらの三つ組に対して，(1) 毎日新聞 8 年分のコーパスからの検索，(2) 出現パタンの抽出，(3) 対象物，属性名，属性値のそれぞれに対応する語の変数化，を行ない 表 1 に示すような 34 個の抽出パタンを得た．それぞれの抽出パタンは形態素単位の依存構造となっている．変数には品詞が名詞の下位クラスであるという制約のみを加えた．

表 1: 獲得したパタンの例

パタン	パタンを生成したシードの出現文脈
OのA、Vが O（AV）を VOAは Oを退団するVA	ミュージカル の 本場、ブロードウェイ が 鳥海山（標高 2236メートル）を 久米豊 日本自動車工業会 会長 は 中日 を退団する 西本聖 投手

4.3 スコアリング

今回の実験において 式 (1) で用いる $S_{OA}(O, A)$, $S_{VA}(V, A)$ には，共起パタンにおける共起頻度を用いた．共起パタンの作成には 4.2 節と同じシードを用い基本的に同様の手法を取ったが，三つ組での出現頻度は少ないと予想されたので (対象物–属性名) と (属性値–属性名) の二つ組によるパタンを抽出した．その結果それぞれ 954，221 個の共起パタンを得た．共起尺度には式 (2) で求められる重み付き相互情報量による共起尺度を用いた．

$$weightted\_MI(x, a) = p(x, a) \log_2 \frac{p(x, a)}{p(x), p(a)} \qquad (2)$$

ここで x は対象物または属性値，a は属性名である．$p(x), p(a)$ は，共起パタンが適用されたインスタンス集合における x, a の出現確率を表し，$p(x, a)$ は x と a それぞれの共起確率を表す．毎日新聞 8 年分 (92 年〜99 年) において共起を求めたが，単語の共起だけではデータの量が足りず信頼できる統計量が得られなくなることが予想されたので，共起確率 $p(x, a)$ については対象物や属性値の意味クラスによるスムージングを行なうことによりこの問題の解決を図った．たとえば「鹿児島大学」と「学長」の共起頻度は低いかもしれないが，もし「鹿児島大学」が<学術機関>という意味クラスに属することが分かれば，<学術機関>と「学長」の間の共起情報を用いてスムージングすることにより統計量の信頼性を上げることができる．意味クラスには IREX の 8 種類の NE のクラスを用い，式 (3) によりスムージングを行なった．そして $p(x, a)$ の代わりに $p'(x, a)$ を用いて 式 (2) を計算した．$p(x_c, a)$ は x のクラスと a との共起確率を表す．x が NE でなかった場合，つまり x_c がない場合は第二項を 0 として計算した．

$$p'(x, a) \quad = \quad \alpha p(x, a) + (1 - \alpha) p(x_c, a) p(x|x_c) \quad (3)$$

5 実験結果

5.1 フィルタリングの結果

提案手法を毎日新聞半年分 (92 年 1 月〜6 月:約 44 万文) に適用し抽出実験を行なった．共起情報は毎日新聞 8 年分から計算した．実験の結果，抽出パタンは 13,743 文にマッチし 19,620 個の属性関係の候補が得られた．この結果から 500 個の候補をサンプリングし人手で調査した結果，216 組の候補が属性関係を持っていた．この候補を提案手法によりフィルタリングした結果を表 2 に示す．表 2 の (i) はパタンによって抽出された三つ

表 2: パタンによる抽出結果

	精度	再現率	F 値
(i)	0.432(216/500)	1.000(216/216)	0.603
(ii)	0.512(208/406)	0.963(208/216)	0.669
(iii)	0.565(208/368)	0.963(208/216)	0.712

組のフィルタリング前の結果を示している．したがって精度の分母はパタンが出力した候補数 (500)，分子はパタンが発見した正解の数 (216) である．この実験では提案手法のフィルタリングの効果を調べることが目的なので，パタンにより抽出された属性関係候補中の属性関係の総数を再現率の分母とした．したがって (i) の再現率は 1 となっている．対象とする文書集合全体に対する再現率については 5.2 節で議論する．

(ii) はフィルタリングの結果である．提案手法では，ランキングにおける閾値 N とスムージングの係数 α がパラメータとなっている．これらの値を N は 1〜10 (step 1)，α は 0〜1 (step 0.1) のそれぞれに設定し実験した．その結果 $N = 5, \alpha = 0.7$ のときに最も良い結果が得られ，フィルタリング前に比べ F 値を 6.6 ポイント (10.9%) 向上させることができた．再現率は上述の意味での正解総数におけるフィルタリングを通過した正解の割合である．

実験結果の例を示す．パタンにより (1) のように三つ組候補が抽出されているときに，推定した上位 5 位 (表 3) の中にパタンで抽出された属性名「書記」が含まれているので，この候補はフィルタリングを通過する．また (2) のようにパタンが適用されているとき，パタンによって抽出された「在外」は対象物と属性値から推定した属性名の上位 5 位 (表 3) には現れないので，この候補は棄却される．

(1) 米朝は $\overset{object}{\text{朝鮮労働党}}$ の $\overset{value}{\text{金容淳}}$ $\overset{attribute}{\text{書記}}$ が一月訪米、...

(2) $\overset{object}{\text{ソ連}}$ の $\overset{attribute}{\text{在外}}$ $\overset{value}{\text{資産}}$ に関する協定...

表 2 の (ii) はパタンのみに比べると精度が向上しているが，約半数もの事例で属性関係でない事例を属性関係として抽出している．精度を低くした原因を調査した結果，原因の一つに文書中の三つ組が複数のパタンにマッチしたことが挙げられた．そこである三つ組が複数の抽出パタンにマッチし属性関係のラベル (対象物，属性名，属性値) が異なって推定されていた場合はスコアの高い方だけを選択し低い方は棄却するようにしたところ，表 2 の (iii) に示すように精度を上げることができた．

表 3: ランキング例

(1)			(2)		
1	**書記**	**2.89e-08**	1	問題	2.74e-10
2	団長	2.89e-10	2	会議	1.34e-10
3	総書記	2.79e-10	3	国家	8.39e-11
4	委員長	2.49e-10	4	外相	4.30e-11
5	主席	1.73e-10	5	大使	3.93e-11

精度を落したその他の原因は，属性関係を持たない三つ組の要素が文書中で多く共起するために高いスコアを得たことであった．(3) がその例である．

(3) ワシントン発の聯合通信によると、$\overset{value}{\text{米}}$ $\overset{object}{\text{国防総省}}$ $\overset{attribute}{\text{スポークスマン}}$ は２日、...

このパタンは (4) のように正しく適用される場合も多いので，パタンや共起尺度以外の情報によりこれらを区別できなければならない．

(4) $\overset{value}{\text{ブッシュ}}$ $\overset{object}{\text{米}}$ $\overset{attribute}{\text{大統領}}$ は三十一日

また再現率を低くした原因としては，重み付き相互情報量を使っているために語単体での出現頻度が小さい属性名に対しては小さい値しか与えられず，そのためそのような属性名が選ばれなかったことが挙げられる．この問題への最も簡単な解決策としてはより大量の文書データで統計的共起尺度を計算することが考えられる．しかし特に固有表現については出現回数も限られているため，文書データを増やすだけでなく，スムージングの改良がより重要である．今回の実験では IREX の 8 種類だけを用いてスムージングしたが，より精緻な意味クラスを用いることでより正確なスムージングができるだろう．

5.2 再現率の調査

5.1 節で示した再現率は，フィルタリングの効果を調べるためのものだったので，パタンがマッチした個所に含まれる属性関係の数に対する獲得できた属性関係の割り合いを示していた．本節では対象文書中の属性関係の数に対する抽出できた属性関係の割り合いを調査する．

今回の実験で抽出に用いた属性リスト (1038 種類) を抽出対象テキスト (毎日新聞 92 年 1 月〜6 月) から検索した結果，出現個所は 285,671 だった．一文に複数の属性名が出現する場合は，それらを個別に出力している．この結果から 500 事例をサンプリングし，属性関係を含むかを人手で調査した．その結果 66 事例において属性関係が見付かった．したがって 285,671 の約 13.2%(66/500) である 約 37,000 の属性関係が対象文書中に存在すると大雑把に見積れる．一方今回の抽出実験では 19,620 の候補から 500 の候補をサンプリング

して調査した結果，そのサンプル中に 216 の属性関係が含まれていた．サンプル中の約 43%(216/500) に属性関係が含まれていたので，パタンで抽出した 19,620 の候補中には約 8,500 の属性関係があったと見積れる．したがって，今回の実験における対象データに対する再現率の上限は約 23%(8,500/37,000) となる．このように再現率の上限が低くなった原因を調査した結果，問題点は (a) パタンでは捉えきれない複雑な形で三つ組が出現していた，(b) 照応・省略のために，三つ組をパタンでとらえられない，(c) 対象物を NE に限定していた，(d) 依存構造での照合の失敗，の 4 つにまとめられた．例文 (5) , (6) は (a) の例である．この例では「大統領」と「アフガニスタン」の関係や「貴花田」と「父親」の関係をとらえる必要がある．このような事例に対しては照応・省略解析とともにテキストの深い解析が必要である．

(5) アフガニスタン^{object} のカブール政権は十八日夜、ナジブラ^{value} 大統領^{attribute} の失脚に伴い、...

(6) 貴花田^{object} の姿勢、だれに似ているかと思いをめぐらすと、やはり人気力士だった貴ノ花^{value}、つまり 父親^{attribute} とダブる。

6 考察

本稿で提案したスコアは対象物と属性値が与えられたときに属性らしさを相対的に評価するだけである．つまり三つ組の候補が属性関係を持っていた場合には高いスコアが得られるが，属性関係を持たないときにそのことを積極的には示さない．そのため高い精度が得られないという結果となった．属性関係を持ちうるかの判定を行うために三つ組のスコアの絶対的な値を使うことも考えられたが，対象物–属性値対ごとにスコアはばらついており絶対的な閾値は決められなかった．今回は S_{OA}, S_{VA} の計算に単純な共起情報だけを用いたが，スコアリングの手法には確率的なモデルを用いる手法や，すでに知っている三つ組との類似度を用いる，または対象物またはそのクラス毎に持ちうる属性の知識を収集しておき属性関係の抽出に用いるなど，この他にも改良する余地がある．

本タスクではドメインを限定しないという点においてこれまで研究が進められてきた MUC 形式の IE を拡張した．そのために一般化したパタンを用いたが，パタンを使うというアプローチ自体は変わっていない．その限界が 5.2 節で示した調査結果に表れている．より高い再現率を得るためには，より深い文 (文章) の解析を伴う処理が必要となる．

7 まとめ

本研究では，ドメインを指定せずに（対象物，属性名，属性値）をテキストから抽出するというタスクを設定し，抽象化したパタンと統計量を組み合せて用いる手法により，パタンのみを用いた場合に対して F 値を約 6.6 ポイント向上させられることを示した．さらに精度を上げるために，共起情報だけでなく対象物またはそのクラスの持ちうる属性の知識を収集し，属性関係の抽出に用いることが有効だと考えられる．また対象テキスト全体に対して再現率を調べた結果，本手法の再現率の上限は約 23% という結果となった．再現率を上げるためには 5.2 節で述べたようにテキストの深い解析・理解が必要となる．

参 考 文 献

[1] Sergey Brin. Extracting patterns and relations from the world wide web. In *WebDB Workshop at 6th International Conference on Extending Data base Technology, EDBT'98*, 1998.

[2] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. Web question answering: Is more always better? In *SIGIR 2002: 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 291–298, 2002.

[3] Michael Fleischman, Eduard Hovy, and Abdessamad Echihabi. Offline strategies for online question answering: Answering questions before they are asked. In *41st Annual Meeting of the Association for Computational Linguistics*, 2003.

[4] Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. Question answering challenge for five ranked answers and list answers - an overview of NTCIR4 QAC2 subtask 1 and 2. In *Working Notes of the Third NTCIR Workshop Meeting: QAC2*, 2004.

[5] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In *42th Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, 2004.

[6] IREX 実行委員会（編）. *IREX*, 1999.

[7] Cody C.T. Kwok, Oren Etzioni, and Daniel S. Weld. Scaling question answering to the web. In *the Tenth International World Wide Web Conference (WWW10)*, 2001.

[8] Jimmy Lin and Boris Katz. Question answering from the web using knowledge annotation and knowledge mining techniques. In *Twelfth International Conference on Information and Knowledge Management (CIKM 2003)*, 2003.

[9] Ellen Riloff and Rosie Jones. Learning dictionaries for information extraction by multi-level boot strapping. In *Sixteenth National Conference on Artificial Intelligence*, 1999.

[10] 高橋哲朗, 乾健太郎, 松本裕治. テキストから属性関係を抽出する. 情報処理学会自然言語処理研究会 (NL-164), pp. 19–24, 2004.

[11] 高橋哲朗, 乾健太郎, 関根聡, 松本裕治. 質問応答に必要な言い換えの分析. 言語処理学会 第 10 回年次大会, 2004.

[12] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 71–78, Philadelphia, July 2002. Association for Computational Linguistics.