2004－NL－164（4）
2004／11／4

テキストから属性関係を抽出する

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 情報工学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
E-mail: †{tetsu-ta,inui,matsu}@is.naist.jp

あらまし　質問応答で用いる知識をオフラインで獲得するために (対象物，属性名，属性値) という三つ組の属性関係をテキスト中から抽出するタスクを考え，抽象化したパタンによりドメインを限定せずに三つ組の候補を抽出し，統計量を使ってそれらをフィルタリングする手法を提案した．実験の結果，この手法を用いることによりパタンのみで抽出した場合に比べて高い精度で属性関係の三つ組を抽出できることを示せた．本研究ではまた対象物と属性値だけをパタンにより抽出し，それらの間の関係の推定を行なった．その結果，既知の属性については高い精度で属性名を推定できることを明らかにした．

キーワード　属性，情報抽出，質問応答，関係獲得

Automatic Extraction of Attribute Relations from Text

TAKAHASHI TETSURO†, INUI KENTARO†, and MATSUMOTO YUJI†

† Graduate School of Information Science, Nara Institute of Science and Technology
Takayama, Ikoma, Nara 630-0192, JAPAN
E-mail: †{tetsu-ta,inui,matsu}@is.naist.jp

Abstract　This paper describes an algorithm to extract triplets of attribute-value relations (Object, Attribute, Value) from text for the purpose of building a knowledge base for question answering. In our approach, the system first extracts triplets using abstracted surface language patterns, and then use a statistical model to filter out implausible ones. Our experiment showed that the method performed well in both recall and precision.

Key words　attribute, information extraction, question answering, relation acquisition

1. 目　　的

本研究の目的は，テキストから（対象物，属性名，属性値）の三つ組を網羅的に抽出することである．本研究では，対象物，属性名，属性値をそれぞれ以下のように定義した．

対象物　特定の物体
属性名　対象物の特徴・性質を現わす関係
属性値　特定の対象物に対して与えられた属性名の関係により一意に決まる値

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．

本タスクにより抽出された三つ組は質問応答のためのデータ構築に有用である．質問応答におけるデータの使い方には，オンザフライでテキストデータから回答を探す方法と，オフラインであらかじめテキストデータからデータベースを作っておき，そこから回答を探す方法がある．Lin ら [1] はユーザの質問の種類の数と聞かれる回数は Zipf's Law に従うという調査結果を示している．すなわち頻繁に聞かれる質問ほどその種類は少なく，稀に聞かれる質問の種類は多い．"What is the population of x?" のように頻繁に聞かれる質問は，特定の限られた種類の関係に関する質問であるため，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．Lin らは構造化された Web 上の知識からデータベースを静的に作成し Web のオンライン検索とともに質問応答に使っている．また Fleischman ら [2] は"Who is X"のタイプの質問に対して，オフラインで事前に知識を獲得し，質問応答に用いるアプローチをとっており，同種の質問に関して精度が向上したことを報告している．

筆者らが QAC2 [3] の質問集合について調査した結果 [4]，24% が (1a) に示すように属性を尋ねる質問だった．また (1b) は，言い換えると (1c) のように属性を尋ねている質問と見ることもできる．このようなケースを含めると属性を尋ねる質問の割合は全体の 40% であった．

(1) a. 石ノ森章太郎さんの出身地はどこですか
b. アントニオ猪木の引退試合はどこで行われましたか
c. アントニオ猪木の引退試合の場所

2. 先行研究

2.1 MUCにおける情報抽出

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，以下の相異点がある．

- シナリオを限定しない
- 属性関係にある三つ組のみを抽出する

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．たとえば，"vehicle launch events"というシナリオにおいて，"launch_date"，"launch_site"，"vehicle_info"など，いくつかの決められた要素を抽出する．このようにシナリオが限定されている場合は，そのシナリオに特化したパタンや辞書の作成または獲得が可能であるが，今回のタスクではシナリオを限定しないためそのようなアプローチをとることができない．

Brin [5] やRiloffら [6] はドメイン固有の辞書項目とそれが出現するパタンを交互に抽出する手法を提案している．Brinの手法では本の著者とタイトルの関係を抽出しており，それらの関係を表すさまざまな文脈をパタンとして獲得できる．またそのパタンにより著者とタイトルの関係にある新しい組を見付けることができる．しかしこの手法は，対象物と属性値を固定して属性名を獲得する，または属性名を固定して対象物と属性値を獲得するものであり，特定の関係にある組のみを抽出する枠組となっている．そのため学習データのドメインに依存してしまい，今回のタスクで求められている多様な属性関係を抽出することはできない．それに対し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．この点でMUCの問題設定の一般化となっており問題の範囲を広げている．

Khooら [7] は人手で作成したパタンを用いてMedlineから*Cause-Effect*の関係の抽出を行なっている．MUCにおけるIEが特定のシナリオに焦点を当てていたのに対し，彼らは特定の関係に焦点を当て，対象をあらゆるシナリオに広げている．Khooらのタスクでは*Cause-Effect*という関係にある2つの値だけを抽出していたのに対し，本タスクでは属性関係というメタな関係の中で，その関係の種類も属性名として抽出する．

Hasegawaら [8] 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BrinやRavichandranら [9] と同様にNEの対から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共起するものでないと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，対象物と属性値が直接共起しない場合でも属性関係を抽出できる枠組を提案する．

MUCのサブタスクにおいても関係の抽出が行なわれていた．MUC-6に始まったTemplate Element（TE）は，ある名詞についてその名詞が持つname, type, descriptor, categoryの属性値を求めるタスクである．またMUC-7で行なわれたTemplate Relation（TR）は名詞間の関係を求めるタスクであり，MUC-7では，organizationにおけるemployee_of, product_of, location_ofの3つの関係を抽出するタスクであった．これらのタスクが数個の属性関係を抽出するものであったのに対し，本研究ではより多様な関係の抽出を目指す．

MUCではシナリオ毎に取りうる情報は異なっていた．本タスクではシナリオを限定しないが，あらゆるシナリオについてそのシナリオが取りうる情報を抽出するのは困難である．そこで今回はシナリオを限定しない代わりに，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する．

2.2 機械学習を用いた情報抽出

提案手法ではパタンと統計的な手法を用いるが，異なるアプローチとして機械学習を用いる手法も考えられる．Zelenkoら [10] は機械学習により*Person-Affiliation*, *Organization-Location*などの関係の抽出を行なった．また抽出に直接機械学習を用いるのではなく，あらかじめ人が用意したパタンを用いて文書から候補を大まかに抽出し，教師あり学習を用いてそれらをフィルタリングするという手法も提案されており [2]，彼らの実験においてそれぞれの手法はある程度うまく働くことが示されている．

しかしこの手法を今回のタスクに適用するためには，関係(注1)毎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，トレーニングデータの作成に大きなコストが必要となるので本タスクに機械学習を適用するのは難しいと言える．

3. 提案手法

多様な種類の属性関係を獲得するためには，属性名毎にトレーニングデータを用意しパタンと辞書項目を獲得するという方法が考えられる．しかし属性名の種類が多くなるとそれぞれについてトレーニングデータを作成することは困難となるため，今回の問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

- 抽象化したパタンを使う
- 抽象化した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する

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．
（2） あらゆる属性名において，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．

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

(注1)：本稿で言う属性名

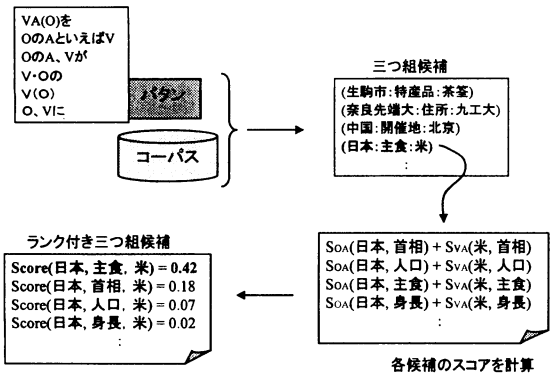

図 1　アプローチの概要

ここで S_{OA}, S_{VA} はそれぞれ

$S_{OA}(O, A)$: $O(bject)$ が $A(ttribute)$ を属性としてどれくらい持ちやすいか

$S_{VA}(V, A)$: $V(alue)$ 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}(日本, 主食) は「日本」がどれくらい「主食」という属性を持ちうるかを表し，S_{VA}(米，主食) は，「米」がどれくらい「主食」の値となりうるかを表す．

本タスクでは三つ組を獲得することが目的であるが，三つ組がそろって出現する確率は低いと考えられる．そこで提案手法では，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している．

この手法の利点の一つに，対象物と属性値だけをとるパタンにも適用できることが挙げられる．実文書の中には (2a) のように属性名が記述される場合だけでなく (2b) のように属性名が陽には記述されない場合が多くある．

(2) a. $\overset{object}{\text{大竹}}$ の $\overset{attribute}{\text{身長}}$ は $\overset{value}{\text{208cm}}$

b. $\overset{object}{\text{大竹}}$ は $\overset{value}{\text{208cm}}$

今回提案したアルゴリズムはこのような事例に対して属性名の推定を行なうことができる．

4. 実　　験

4.1 パタンの作成

まず人手により約 70 個の三つ組をシードとして用意した．そして形態素単位の係り受け木に変換したコーパスから，シードの三つ組に含まれる語をすべて含む最小の部分木を抽出し，その中の対象物，属性名，属性値に対応する語をそれぞれ変数化し，属性名付きパタンを作成した．変数には，品詞が名詞でなければならないという条件のみを加えた．また属性名無しパタンは対象物と属性値を含む部分木から同様に作成した．この手法により毎日新聞 8 年分のコーパスから表 1 に示すような属性名有りパタン 34 個，属性名無しパタン 341 個を得た．

表 1 の属性名有りパタン 1 は実際には以下のように記述されている．

```
1: M(ID:1& M:2& P:名詞& W:O)
```

表 1　獲得したパタンの例

属性名有りパタン	属性名無しパタン
1: OのA、Vが	1: O、Vに
2: V（OA）が	2: V・Oの
3: OのAであるV	3: V（O）

```
M(ID:2& M:3& P:助詞-連体化& W:の)
M(ID:3& M:4& P:名詞& W: A)
M(ID:4& M:7& P:記号-読点& W:、)
M(ID:7&M:8& P:名詞& W:V)
M(ID:8& M:9& P:助詞-格助詞-一般& W: が).
```

M,P,W はそれぞれ係り先 ID，品詞，表層形である．

4.2 Base NP chunking

4.1 節で示したパタンは形態素単位で適用する．そのためこのパタンをそのまま (3) の文に適用すると，この例のように接尾辞のみが抽出されてしまう．

(3) a. 伊藤隆男 $\overset{value}{\text{氏}}$（$\overset{object}{\text{資生堂}}$ $\overset{attribute}{\text{社長}}$）が...

b. キューバ政府スポーツ $\overset{object}{\text{庁}}$ $\overset{attribute}{\text{長官}}$ である $\overset{value}{\text{マルティネス}}$ 選手は..

この問題を防ぐために，まず対象コーパスに対して Base NP chunking を行い，その後パタンを適用した．これにより (3) 中の名詞句もパタン中の変数とマッチできるようになる．

Kudo ら [11] が提案するように教師あり学習を用いることにより高い精度で Base NP chunking を行なうことができるが，学習事例の作成にかかるコストの問題から，今回はルールベースで Base NP chunking を行なった．使用した規則を以下に示す．

- カタカナ連続 → カタカナ
- 接頭辞 + NE → NE
- 接頭辞（接頭詞-名詞接続） + 名詞 → 名詞
- 名詞 + 接尾辞（名詞-接尾） → 名詞

これらの規則によりカタカナ語や接頭辞，接尾辞に関する問題は解決されたが，名詞の連続からなる複合語のチャンキングができなかった．たとえばこれだけの規則では，「マリアナ + 海溝 → マリアナ海溝」のようなチャンキングができない．“NE + 名詞 → NE” のような規則があれば「マリアナ海溝」のチャンキングも可能となるが，この規則は「北朝鮮 + 内」，「米 + ソ」のような誤ったチャンキングを行なってしまうので使用していない．

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，どちらも重み付き相互情報量による共起尺度を用いた．$S_{OA}(O, A)$ と $S_{VA}(V, A)$ のそれぞれの定義に対して重み付き相互情報量は粗い近似ではあるが，今回の実験ではその近似でどれだけの精度が得られるかを確認する．

スコアリングの手法については 6.2 節で詳しく議論する．

4.4 対象物の限定

今回の実験では対象物を固有表現に限定する．最終的な目的としては普通名詞も対象物として扱いたいが，普通名詞は指示的に使われている場合に対象物の特定が困難であるという問題

がある．たとえば (4) から (5a) の情報を抽出できたとしても対象物を特定できていないので有用な情報にはならない．また (5b) のように対象物を特定できれば有用な情報になる可能性はあるが，このような対象物の特定は容易にはできない．

(4) 太郎は昨日，定価 3,000 円の CD を買った．

(5) a. (CD$_{object}$, 定価$_{attribute}$, 3,000 円$_{value}$)

b. (太郎が昨日買った CD$_{object}$, 定価$_{attribute}$, 3,000 円$_{value}$)

普通名詞が (6) の「CD」のように総称的な使われ方をしている場合には，ここから一般的な知識を抽出できる．しかし，普通名詞が総称的であるか指示的であるかの区別は現在の技術では容易ではない．

(6) CD$_{object}$ の 記憶容量$_{attribute}$ は 650M バイト$_{value}$

以上の理由から，今回の実験では対象物を固有表現に限定した．

4.5 属性名の限定

情報抽出では抽出すべき属性名がタスクの中で与えられていたので，属性とは何かという議論をする必要はなかった．それに対し本タスクでは属性名も抽出の対象となり，どこまでを属性と考えるかという問題が生じる．たとえば「出身地」,「社長」は特定の値を持つ属性名だと考えられるが,「話」,「願望」は特定の値で表わせない場合もあり属性名としては適さない．3 節で述べたアルゴリズムでは本来あらゆる属性名について三つ組のスコアを計算することが望ましいが，上に挙げた理由から本研究では属性名のリストを与えることで属性名を外延的に定義する．

属性名のリストを作成するために，今回は「*NE* の x」というパタンの x に入る表現をテキストコーパスから抽出し，そこから人手で属性名となりうる語を選別した．*NE* には IREX [12] で定義された 8 種類のクラスの中から，*ORGANIZATION*, *PERSON*, *LOCATION*, *ARTIFACT* の 4 種類を用いた．テキストコーパスには毎日新聞 8 年分，NE チャンカには CaboCha [13] を用いて抽出を行ない，1038 種類の属性名からなる属性リストを作成した．

4.6 デ ー タ

上記のアルゴリズムを毎日新聞半年分に適用し，属性名付きパタン，属性名無しパタンのそれぞれを用いて抽出実験をした．共起情報は毎日新聞 8 年分から計算した．

5. 実験結果

5.1 属性名付きパタン

属性名付きパタンは 266 文にマッチし 266 個の属性関係の候補が得られた．これらを人手で調査した結果，属性関係を持つ候補数は 258，持たない候補数は 8 であった．またパタンにより正しく属性関係を抽出できた数は 219 であった．

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「パタンのみ」はパタンによって抽出された三つ組のフィルタリング前の精度と再現率を示している．したがって精度の分母はパタンが出力した候補数 (266)，分子はパタンが発見した正解の数 (219) である．再現率は本来対象とした全文書中に含まれる正解の総数を分母として計算すべきであるが，

表 2 属性名付きパタンによる抽出結果

	パタンのみ	フィルタリング
精度	219/266(0.82)	216/238(0.91)
再現率	219/258(0.85)	216/258(0.84)
F 値	0.84	0.87

今回の実験では提案手法であるフィルタリングの効果を調べることが目的だったので，パタンが適用された文に含まれる属性関係の総数を再現率の分母とした．この再現率 (0.85) はフィルタリングの結果得られる再現率の上限となっている．「フィルタリング」の精度はフィルタリングを通過した候補 (238) における正解の割合であり，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．

今回の実験では N の値は 10 とした．つまりパタンが見つけた属性名の候補が，提案手法により推定しランキングした属性名の上位 10 位以内に入っているときに，属性名であると判断した．このフィルタリングにより，再現率を 1% の降下に抑えながら精度を約 10% 上昇させることができた．再現率を下げた 3 事例 (219 − 216) はいずれも十分な統計量を得ることができなかったために，スコアを求められなかったことが原因であった．

実験結果の例を示す．パタンにより (7) 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 10 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．

(7) 米朝は 朝鮮労働党$_{object}$ の 金容淳$_{value}$ 書記$_{attribute}$ が一月訪米、米...

表 3 正しい属性名の選択の例

1	**書記**	**2.89e-08**	6	担当	1.62e-10
2	団長	2.89e-10	7	問題	1.25e-10
3	総書記	2.79e-10	8	代表団	1.06e-10
4	委員長	2.49e-10	9	会議	6.32e-11
5	主席	1.73e-10	10	部長	3.09e-11

また (8) のように候補が抽出されているとき，パタンによって抽出された「事務」は，対象物と属性値から推定した属性名の上位 10 位 (表 4) には現れないので，この候補は棄却される．

(8) 長田譲$_{value}$ ・ 大阪宗教者平和協議会$_{object}$ 事務$_{attribute}$ 局長は戦死者...

表 4 誤った属性名の棄却の例

1	局長	1.11e-06	6	顕官	0.0
2	チーフ	0.0	7	通称	0.0
3	市長	0.0	8	総督	0.0
4	首謀者	0.0	9	首席	0.0
5	地頭	0.0	10	小吏	0.0

5.2 属性名無しパタン

属性無しパタンを用いることで対象物–属性値の対が獲得できる．ここではこの対から属性名の推定がどれくらいできるかを調査する．

属性名無しパタンから得られた 521 個の対象物–属性値を人手で調べた結果，それらが属性関係を持つ事例数は 292，持たない事例数は 229 だった．また，属性関係を持つ 292 事例のう

ち，207事例にはテキスト中に属性名が現れていたが，85事例において属性名が現れていなかった．

テキスト内に属性名を持つ207事例のうち，本手法では204事例を上位10位以内に推定できた．それらの順序を表5に示す．推定できなかった3事例 (207 − 204) のうち2事例では文

表5　テキスト中に現れる属性名の推定

順位	1	2	3	4	10	計
事例数	137	55	8	2	2	204

書中に出現する属性名が属性名リストに入っていなかったことが原因だった．また十分な統計量が得られなかったために共起を計算できなかったことが原因となった事例が1事例あった．

テキスト内に属性名が現れない85事例のうち，17事例は1位で，2事例は2位で推定できた．たとえば (9) のように対象物と属性値が抽出されたときに，表6に示す結果が得られた．ここでは本文には現われていない「選手」という語を属性名として推定できている．

(9)　前回大会で 森下広一$_{value}$ (旭化成$_{object}$) が出した昨年世界最高で大会記録の2時間8分53秒を...

表6　属性名の推定

1	選手	**8.92e-10**	6	兄弟	5.27e-12
2	代表	8.44e-11	7	アンカー	3.82e-12
3	最高記録	3.47e-11	8	ゴール	1.66e-12
4	記録	3.28e-11	9	エース	1.03e-12
5	ランナー	8.63e-12	10	引退	6.89e-13

テキスト中に属性名が現れない事例のうち，66事例において属性名を推定できなかった．これには次のような課題設定自体の問題もある．たとえば (10) のように，考えられる属性名が「場所」や「国籍」のように抽象的である場合一意に決めるのは困難である．

(10) a.　米$_{value}$・カリフォルニア大学$_{object}$ が数十年前に...

b.　韓国$_{value}$ の ＫＢＳテレビ$_{object}$ は八日、...

提案手法では対象物–属性値の対から属性名の推定を行なうだけであり，属性関係を持たないという判別ができない．属性名の持つスコアによりこの判別の可能性が期待できたが，対象物–属性値対毎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．この点において提案手法は拡張が必要である．

6. 考　　察

6.1　チャンキング誤り

5.1節の (8) ではパタンにより正しい属性名「事務局長」を抽出することができず，その一部である「事務」だけが抽出されていた．この問題の原因は「事務局長」という複合語のチャンキングととらえることができるので，その対策としては，4.2節で行なった Base NP chunking の精度を上げ，パタンを適用する前にあらゆる複合語のチャンキングを済ませておくことが考えられる．しかし名詞句の範囲を曖昧性なく決めることは容易ではない．たとえば (11) において，「平野貞夫」を属性値とする対象物と属性名は，我々の定義において (12) の4種類が考えられる．

(11)　自民推薦の無所属、平野貞夫・衆院事務局委員部長が既に出馬表明している。

(12) a.　(衆院事務局委員部^{object}, 長^{attribute})

b.　(衆院事務局委員^{object}, 部長^{attribute})

c.　(衆院事務局^{object}, 委員部長^{attribute})

d.　(衆院^{object}, 事務局委員部長^{attribute})

上記の理由から我々は，パタンの適用や共起の計算において，複合語としてとりうる可能性をすべて抽出候補として扱えるよう提案手法を拡張することを考えている．

6.2　スコアリング

今回は三つ組候補のスコア計算に一文単位での共起を用いたが，一文では直接関係を認めるには広すぎるため，ウインドウ内における共起，あるいは係り受け関係における共起を用いることも考えられる．また本研究では「NE の x」というパタンにより属性名をあらかじめ抽出したが，このパタンは対象物と属性名の関係を表わすパタンであるので，これに加え属性値と属性名の関係を表わすパタンを用いることで，より正確な共起情報を得られる．たとえば一文単位で共起を計算した場合，(13) のような文から（日本，首相）や（アメリカ，大統領）という本来抽出すべき正しい共起情報だけでなく，(日本，大統領）や（アメリカ，首相）という誤った共起情報も同じように扱われてしまう．しかし「NE の x」のようなパタンにおける共起だけを考えることにより，上記の正しい関係だけをカウントできる可能性が高くなる．

(13)　日本の首相にとってアメリカの大統領はだれであっても同じということか。

今回用いた三つ組候補のスコアは頻度を元に計算しているため，頻度が低いデータについてはスコアの信頼性がなくなる．この問題には，対象物や属性値の意味クラスを定義し，意味クラスと属性名間の共起を用いることで対処できる可能性がある．たとえば「奈良先端大」と「学長」の共起頻度は低いかもしれないが，もし「奈良先端大」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」の間の共起情報でバックオフすることにより統計量の信頼性を上げることができる．今回の実験では数値表現についてのみ単位を用いた抽象化を行なったが，他の種類の語についても意味クラスによるバックオフを試す必要がある．

今回は S_{OA}, S_{VA} の計算に単純な共起情報だけを用い，それだけでどれくらいの精度が得られるのかを調査した．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，すでに分かっている三つ組との類似度を用いる手法など，この他にも考える余地がある．

6.3　属性関係に関する知識

6.3.1　属性名と対象物・属性値の間の知識

対象物と属性値にクラスを定義できれば，クラス毎にどのよ

表 7 NE を用いたフィルタリング

	pattern のみ	NE フィルタリング
精度	219/266(0.82)	129/132(0.98)
再現率	219/258(0.85)	129/258(0.50)
F 値	0.84	0.66

うな属性を持つかという知識を与えることができる．この知識の効果を見るために，今回用いた属性名リストに，

- その属性名を持ちうる対象物のクラス
- その属性名の値となりうる属性値のクラス

を人手で記述し，その知識を三つ組候補のフィルタリングに用いた．クラスには IREX で定義された NE クラスを用いた．

実験の結果を表 7 に示す．この結果から．上記の知識を用いることにより，再現率は低いが高い精度で三つ組を抽出できることが分かる．このような知識を網羅的に用意することは困難であるが，このような知識が部分的にでもあったときに，それを活かせるような枠組に提案手法を拡張していきたい．

この実験では IREX で定義されたクラスのみを用いたが，より詳細なクラスを定義し正確にクラスを当てることができれば，より高い精度で属性関係にある三つ組を抽出できると考えられる．また今回実験した NE フィルタリングでは対象を NE に限定していたが，普通名詞に関してもシソーラス上のクラスなどを用いることにより同様の制約を記述できる．

Yoshida ら [14] はあらかじめ web から抽出したオントロジーを用いて，web 上の表から属性名とその値を抽出している．本タスクにおいても属性名と属性値の間でオントロジーを用いることで，より抽出の精度を高めることができると考えられる．

6.3.2 属性名の獲得

本研究では対象とする属性名を限定しているため，それら以外の関係を見付けることはできない．今回はパタンを用いてコーパスから属性名を抽出したがそれだけであらゆる属性名を網羅的に獲得するのは困難であり，属性名を発見的に追加できる枠組が必要である．

笹野ら [15] は本研究における「属性名」にあたる語を「名詞の格フレーム」と呼び機械的に収集する手法を提案している．また Hasegawa ら [8] の手法では，前述したように関係の自動抽出も行なう．これらの手法を用いてより多くの属性を得ることができれば，属性関係の三つ組をより多く獲得できる．

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再現率の降下を 1% に抑えながら精度を約 10% 上昇させられることを示した．本研究ではまた，対象物と属性値だけをパタンにより抽出し，それらの間の関係の推定を行なった．その結果，既知の属性名については高い精度で推定できることを明らかにした．

今後は 6 節の議論を基に，より高い精度で抽出できるよう提案手法を拡張していきたい．また今回の実験ではパタンが適用された文書に含まれる正解数を基に再現率を求めたが，全文書に対する再現率についても今後調査する必要がある．また今回は対象物と属性値の関係として「N1 が N2 を書く」というようなイベントの関係は扱わずに，名詞で表わせる属性名の関係のみを扱ったが，今後は属性だけでなくイベントも同様に抽出していきたい．

文 献

[1] J. Lin and B. Katz: "Question answering from the web using knowledge annotation and knowledge mining techniques", Twelfth International Conference on Information and Knowledge Management (CIKM 2003) (2003).

[2] M. Fleischman, E. Hovy and A. Echihabi: "Offline strategies for online question answering: Answering questions before they are asked", 41st Annual Meeting of the Association for Computational Linguistics (2003).

[3] J. Fukumoto, T. Kato and F. Masui: "Question answering challenge for five ranked answers and list answers - an overview of NTCIR4 QAC2 subtask 1 and 2", Working Notes of the Third NTCIR Workshop Meeting: QAC2 (2004).

[4] 高橋, 乾, 関根, 松本："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).

[5] S. Brin: "Extracting patterns and relations from the world wide web", WebDB Workshop at 6th International Conference on Extending Data base Technology, EDBT'98 (1998).

[6] E. Riloff and R. Jones: "Learning dictionaries for information extraction by multi-level boot strapping", Sixteenth National Conference on Artificial Intelligence (1999).

[7] C. S. Khoo, S. Chan and Y. Niu: "Extracting causal knowledge from a medical database using graphical patterns", 38th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (ACL 2000) (2000).

[8] T. Hasegawa, S. Sekine and R. Grishman: "Discovering relations among named entities from large corpora", 42th Annual Meeting of the Association for Computational Linguistics (ACL 2004) (2004).

[9] D. Ravichandran and E. Hovy: "Lerning surface text patterns for a question answering system", 40th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (ACL 2002) (2002).

[10] D. Zelenko, C. Aone and A. Richardella: "Kernel methods for relation extraction", Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, Association for Computational Linguistics, pp. 71–78 (2002).

[11] T. Kudo and Y. Matsumoto: "Chunking with support vector machines", The Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) (2001).

[12] :IREX ワークショップ予稿集 (IREX 実行委員会 (編)) (1999).

[13] T. Kudo and Y. Matsumoto: "Japanese dependency analysis using cascaded chunking", 6th Conference on Natural Language Learning 2002 (CoNLL 2002) (COLING 2002 Post-Conference Workshops), pp. 63–69 (2002).

[14] M. Yoshida: "Extracting attributes and their values from web pages", ACL Student Research Workshop, Philadelphia, Association for Computational Linguistics, pp. 1–6 (2002).

[15] 笹野, 河原, 黒橋："名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析", 言語処理学会 第 10 回年次大会 (2004).