# 意味クラスタリングを用いたWeb文書からの概念階層の取得

三浦 研爾 †　鶴岡 慶雅 †‡　辻井 潤一 †‡
†東京大学　‡CREST, 科学技術振興機構
{miuken, tsuruoka, tsujii}@is.s.u-tokyo.ac.jp

## 1　はじめに

単語の概念階層は自然言語処理における様々なタスクにおいて非常に有用である．このような情報を利用する手段としてWordNet [1], EDR [2] などの巨大な電子辞書を利用する方法があるが，辞書に載っていない意味を扱えないという問題点がある上，これらの辞書は人手で編集されているために内容の維持にコストがかかる．

そのため，構文情報や意味情報のアノテーションがなされていないような普通のテキストを集めたコーパスから概念階層を構築する研究がなされてきた．Hearst [3] らは単純な構文パターンを使い，タグ付けされていないコーパスから概念関係を得る方法を示した．しかし，この方法で得られる関係は多くの正しくない関係 (ノイズ) を含むため，Berland と Charniak [4] はそのようなノイズを除去するための統計的なスコア付け手法を提案した．このスコアは下位概念として現れた単語がある語を上位概念として構文パターンに出現させる条件付き確率を用い，正しくない関係を効率的に取り除くことができる．しかし，この手法は名詞を単純に字面で扱っているために単語の多義性を考慮できず，スコアの値が多義語に対して低くなってしまうという問題がある．

本研究では，前述の多義語のスコアが低くなる問題を解決するための手法を提案する．我々の手法はターゲットとなる多義語の上位概念をクラスタリングすることにより，その語義を自動的に発見する．この手法で使用する名詞階層を得るためのコーパスとしてWeb文書を用いることにより，特定の関係がコーパス上に現れないというスパースネス問題を解決することができる．

## 2　Web文書からの名詞階層の取得

### 2.1　構文パターン

Hearst [3] は概念関係をタグ付けされていない文書から得る方法を示した．例えば，ある語 *NP* の下位概念を得たい場合は，以下のようなパターンを文書から検索する．

“*NP* such as {*NP*,} ... (and | or) *NP*”
“such *NP* as {*NP*,} ... (and | or) *NP*”
“*NP* including {*NP*,} ... (and | or) *NP*”
“*NP* especially *NP*”

例えば，“animals such as”という句で検索をすると下記のような句を含む文書を得られる．

animals *such as* cats, deer, dogs

この句から，animal という単語は cats, deer, dogs の上位概念をあらわす語であるという関係を抽出することができる．

実験においては，名詞句を品詞の単純な正規表現によるパターンマッチで抽出している．

$$NP \simeq \{adj\} * \{noun\} + .$$

### 2.2　Web文書のコーパスとしての利用

このような手法によりドメインに対して高被覆な概念階層を構築するためには，大量のテキストをコーパスとして用意する必要がある．本研究ではWeb文書をコーパスとして利用する.Web文書は多様なジャンルで構成されるテキスト集合としてみることができる．たとえば検索エンジンのGoogleを用いることで, Web上の30億以上の文書にアクセスすることができる．このため，ある語を含む構文パターンがコーパス中に現れない，またはほとんど現れないといったスパースネス問題が軽減され，多くの関係を得ることができる．

### 2.3　スコアづけによる間違った関係の除去

前章に示した手法で得られる関係は多くの正しくない関係（ノイズ）を含むことが知られている．そこで，これらのノイズを除去する処理が必要になる．Berland と Charniak [4] は得られた名詞間関係の信頼性を示すため

に確率統計的なスコア付け手法を提案した．彼らは，語 X が Y の上位概念である信頼性を以下の式で示した．[1]

$$Score(X,Y) = p(X \mid Y) \tag{1}$$

このスコアは語 Y が現れたときに語 X が上位概念としてパターン中に出現する条件付き確率で，X と Y の関係の強さをあらわすと考えることができる．関係の現れた頻度によらないため，コーパス中にあまり現れなかった正しい関係についても高い値を与え，正しい関係として扱うことができる．また，ほとんどのノイズに対してこの確率が低い値をとるために，閾値を設定してフィルタリングを行うことでノイズを効率的に除去できる利点がある．

我々の手法では，$Y$ と $X$ はそのまま名詞 $(W)$ とその上位概念 $(Hypernym(W))$ に対応する．

$$Score(Hypernym(W),W) = p(Hypernym(W) \mid W) \tag{2}$$

予備的な実験から，複数のパターンで観測された関係はより正しい傾向があることがわかった．そこで本研究では，各構文パターン中におけるスコアを足し合わせて新しいスコアとした．

$$Score'(Hypernym(W),W) = \sum_{lsp \in P} Score(Hypernym(W),W,lsp) \tag{3}$$

ここで，P は上位概念を得るために使用する構文パターンの集合である．このようにして計算されたスコアに適切な閾値を設定することによって，ほとんどのノイズを効果的に除去し，精度の高い概念階層を得ることができる．

しかしながら，$W$ が多義語であった場合には正しい関係であっても除去してしまうことがあった．これまでの議論では名詞を文字列として区別しているため，語義ごとに別の語として扱うべき多義語を一つの語として扱っているという問題がある．例えば crane は「鳥」と「建設機械」の 2 つの意味をもつが，式 2 における条件部の crane の生起頻度は両方の意味での頻度の和が使われる．これによって，多義語においては確率的なスコアが本来あるべき値より低くなってしまうという問題が起こる．このような問題に対応するために，多義語を正しく扱えるようにスコアの計算方法を修正する必要がある．

[1] 彼らの研究はコーパスから “全体” と “部分” の関係を得る研究である．実験では X が “全体” であり，Y が “部分” であった．

# 3 多義語の語義発見

前章で示したように，統計的なスコアは多義語に対して低い値をとってしまう．この章では，得た名詞階層から多義語の語義を自動的に発見するアルゴリズムを示す．このアルゴリズムは名詞とその上位概念から，その名詞の語義を得る．

## 3.1 上位概念のクラスタリング

我々の手法では，対象の名詞の上位概念をクラスタリングすることでその語義を得る．ある 2 つの上位概念それぞれについて下位概念として観測された語をベクトルとし，このベクトルに対するコサイン値 [5] を類似度として利用する．この値がある閾値より大きい場合，2 つの上位概念を同じクラスタへ統合する．$W_a$ と $W_b$ を対象の名詞の上位概念とすると，これらの上位概念の集合 $G_a$ と $G_b$ を以下のように記述できる．

$$\begin{aligned} G_a &= \{x \mid x \text{ が } W_a \text{の下位概念 }\} \\ G_b &= \{x \mid x \text{ が } W_b \text{の下位概念 }\} \end{aligned} \tag{4}$$

$W_a$ と $W_b$ について，素性ベクトル $V_a$ と $V_b$ を定義する．ベクトルの各要素 $\{f_i\}$ は $G_a \cup G_b$ の要素のリストとして記述され，以下のように定義される．

$$\begin{aligned} V_x &= \{v_1, v_2, v_3, \ldots v_n\} \ (n = \mid G_a \cup G_b \mid) \\ v_i &= \begin{cases} 1 & (\text{if } f_i \in Hyponym(W_x)) \\ 0 & (\text{otherwise}) \end{cases} \end{aligned} \tag{5}$$

上位概念 $W_a$ と $W_b$ のコサイン値は以下の式で計算される．

$$\cos(W_a, W_b) = \frac{V_a \cdot V_b}{\sqrt{\sum_{a \in V_a} a^2} \sqrt{\sum_{b \in V_b} b^2}}. \tag{6}$$

$W_a$ と $W_b$ は少なくとも 1 つの下位概念（対象の名詞）をもつため，$V_a, V_b \neq 0$ である．実験では $G_a$ と $G_b$ の要素を，2 つ以上のパターンによって観測された下位概念に制限した．

## 3.2 語義クラスタリングアルゴリズム

以下に我々のアルゴリズムを示す．クラスタリングアルゴリズムは single-link の階層クラスタリングに基づく．初期状態では入力された上位概念語それぞれがクラスタを形成する．

まず，入力された上位概念のうち，複合語であるものをそのヘッド（形容詞的な語をすべて取り除いた名詞）に置き換える．この時点で同じヘッドを持つ上位概念は同じクラスタへ統合する．

次にすべてのクラスタの組について，概念階層上の関係がないかを調べる．例えばクラスタ $A$ に属する語 $W_A$ がクラスタ $B$ に属する語 $W_B$ の下位概念であったとき，$A$ に $B$ を統合する．

最後に，残ったクラスタすべてについて以下の 1. 2. を繰り返す．

1. 2 つのクラスタ間のコサイン値を計算する．計算には最初にそのクラスタに含まれていた上位概念を利用する．
2. もしも 1. で計算された類似度が閾値よりも大きければ，2 つのクラスタを 1 つのクラスタに併合する．

得られたクラスタに属する上位概念が, 多義語の一つの語義をあらわす．アルゴリズムはクラスタの類似度を計算するときに各クラスタに対して常に同じ語を用いるため，同じクラスタ内の要素が低い類似度を取ることがあるというチェイニング問題の影響を受けにくいと思われる．

# 4 多義語に対するスコア修正

前章で得られた多義語の語義を利用して，多義語のスコアが低くなる問題を可決する．

我々の提案する新しいスコアでは，多義語 $X$ の意味を区別するために，意味ラベル $S$ を付与する．$(X \rightarrow X_S)$．$X_S$ は “$S$ という意味で使われている $X$” を示す．

次に 得られた名詞間の関係 $\{w, X\}$ を，使われている $X$ の意味によってグループ分けする．$X_S$ をそれぞれの意味に対して別の単語としてみた場合，各 $S$ に対して $X_S$ の上位概念として観測される名詞の条件付確率の総和が 1 にならなければならない．

$$\sum_{w \in Hypernym(X_s)} p(w \mid X_s, lsp) = 1 \qquad (7)$$

この制約に基づいて，提案手法では $p$ の総和を $S$ ごとに正規化する．

$$Score(Hypernym(X_S) \mid X_S)$$

$$= \sum_{lsp \in P} \frac{p(Hypernym(X) \mid X, lsp)}{\sum_{w \in C} p(Hypernym(w) \mid w, lsp)}$$

ここで $C$ は $X_S$ に対する全ての上位概念をあらわす集合である．

修正された計算式では，$X$ の意味の数に応じてスコアが増加する. 特に，出現頻度の低かった意味に対しては修正が強く作用する．

# 5 実験

**名詞階層の取得**　実験では検索エンジンとして Google [6] を用いて文書を取得した．各構文パターンについて検索結果で上位に現れた文書から順に，上位概念を得る場合は 250 文書，下位概念を得る場合は 500 文書を上限として取得した．得られた文書に Brill's tagger [7] を用いて品詞タグを付与し, 名詞間の関係を抽出した．

最初に, Web 文書をコーパスとして使用する効果を調べるために, 語 “country” の下位概念を取得する実験を行った．このときにフィルタリングに用いたスコアの閾値は 0.03 とした．スコアによるノイズ除去を行った結果と，ノイズ除去を行わない場合（ベースライン）を比較して表 1 に示す．ノイズの除去を行った場合, 再現度が落ちているものの正解度で大幅な改善がみられ，結果として F-score が高くなっていることがわかる．

| | 再現度 | 正解度 | F-score |
|---|---|---|---|
| ベースライン | 85.84% | 48.5% | 61.98 |
| ノイズ除去あり | 78.76% | 80.9% | 79.81 |

表 1: “country” の下位概念取得結果.

**多義語の語義発見**　次に上記の実験で得た名詞階層を用いて多義語の語義発見の実験を行った．クラスタリングにおける閾値を 0.05 としたときの, bat の上位概念のクラスタリング結果を表 2 に示す．さらに, クラスタ名 (代表要素) と WordNet の語義を人手で比較した結果を表 3 に示す．実験では WordNet に記述された語義のうち 71% を得ることができた．得られた語義 1 つに対応する WordNet の語義の数は 1.94 個であった．これは我々の手法の, WordNet に対する相対的な粒度と解釈することができる．

| 名詞 | クラスタ名 | 要素 |
|---|---|---|
| bat | mammal | mammal, animal, wildlife<br>bird, pollinator, species |
| | equipment | equipment, softball equipment |

表 2: 語 bat の上位概念のクラスタリング結果.

**多義性の名詞階層への反映**　第 4 章に示したスコアの修正を，表 2 の多義語に対して適用した．

| 名詞 | クラスタ名 | WordNet の語義定義 |
|---|---|---|
| bat | mammal | nocturnal mouselike mammal |
| | equipment | (baseball) a turn batting |
| | | a small racket for playing squash |
| | | a bat used in playing cricket |
| | | a club used for hitting a ball |
| mouse | rodent | any of numerous small rodents |
| | device | a hand-operated electronic device |
| crane | equip-ment | lifts and moves heavy objects |
| | bird | large long-necked wading bird ... |

表 3: 提案手法と WordNet から得られた語義の比較.

| 名詞 (語義) | 上位概念 | 修正後のスコア | 修正前のスコア |
|---|---|---|---|
| bat (animal) | mammal | 0.1363 | 0.0143 |
| | bird | 0.0454 | 0.0047 |
| | wildlife | 0.1363 | 0.0143 |
| | pollinator | 0.1363 | 0.0143 |
| | specie | 0.3181 | 0.0335 |
| | animal | 1.0909 | 0.1148 |
| bat (equipment) | equipment | 0.7500 | 0.0287 |
| | softball equipment | 0.5000 | 0.0191 |

表 4: 語 "bat" の上位概念と修正後のスコア.

表 4 に概念 $bat_{animals}$ と $bat_{equipment}$ に関する上位概念との関係スコアを示す. 提案手法は多義語とその上位概念のスコア, 特に構文パターン中にあまり現れなかった語義に関する上位概念のスコアを高くしている. これによって, 修正前のスコアでは語義 equipment に属する上位概念 (equipment, softball equipment) のスコアが非常に低かったためにフィルタリングで除去されてしまったのが, 本手法によって語義 animal と同等のスコアまで改善されている.

## 6 考察

得られた名詞間関係の頻度統計をみてみると, 関係の中に観測される 2 つの概念は互いに近い関係にある傾向があった. 例えば, 語 bat に対しては, mammal や animal のほうが creature や species といった概念階層の上層に位置する語よりも上位概念として頻繁に観測された. この現象は, これらの構文パターンが概念階層上でより近い階層に位置する概念に対して頻繁に利用されるということを示していると思われる.

我々の語義取得アルゴリズムは最初に複合語をヘッドに置換するため, 結果として得られる語義の粒度が荒くなってしまっている. 例えば語 bat に対する softball equipment という上位概念と cricket equipment という上位概念はこの操作で同じクラスタとして併合されるため, WordNet の対応する語義を別のものとして抽出することができない. この問題はヘッドへの置換を行わないことで解決できるであろう.

## 7 結論

本研究では Web 文書から概念関係を得る方法を示した. 実験では構文パターンを使うことで Web 文書から上位概念と下位概念の関係を得, 得られた名詞階層を利用して多義語の語義を抽出した. 得られた語義は WordNet に記述される語義の 71% をカバーし, 得られた語義 1 つに対する WordNet の語義の数は平均して 1.94 であった. 本研究で提案した新しいスコア手法によって, 多義語に対して確率的スコアが低くなる問題を解決した.

## 参考文献


[1] Christiane D. Fellbaum. Wordnet: an Electronic Lexical Database. In *the Communications of the ACM, 38(11)*, pages 39–41, 1998.

[2] Toshio Yokoi. The EDR Electronic Dictionary. In *Communications of the ACM, 38(11)*, pages 42–44, 1995.

[3] Marti A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *15th International Conference on Computational Linguistics*, pages 539–545, 1992.

[4] Matthew Berland and Eugene Charniak. Finding Parts in Very Large Corpora. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 57–64, 1999.

[5] Gerard Salton and Michael McGill. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983.

[6] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In *the 7th International WWW Conference*, Brisbane, Australia, 1998.

[7] Eric Brill. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT, 1992.