生成語彙論に基づく日本語の特質構造のランキング学習による自動獲得

常吉 高弘　　小町 守　　松本 裕治
奈良先端科学技術大学院大学 情報科学研究科
{takahiro-t, komachi, matsu}@is.naist.jp

1　はじめに

Pustejovsky の生成語彙論 [1] で紹介されている知識表現は，語と語の繋がりからなる句や文の意味解釈や，情報獲得に有用と考えられている．実際に，日本語においては，生成語彙論の知識表現を用いて「家の本」や「私の絨毯」などの名詞句「A の B」の意味解釈をする研究 [2] がなされている．

ところが，現在，大規模な生成語彙論的リソースは公開されておらず，生成語彙論を用いた研究をするためには，まず，人手でリソースを作成する必要がある．そのため，生成語彙論を用いた実用的な研究の数は多くない．

そこで，生成語彙論で用いられる知識表現の中核をなす特質構造を，Web やコーパスから自動獲得する研究が取り組まれている [3, 4, 5]．

[5] の手法は，機械学習を用いて，特質構造を構成する目的役割と主体役割を自動獲得するものであり，パターンマッチングを用いて特質構造を獲得する [3, 4] の手法に対して，獲得パターンを人手で作成する労力が掛からないという利点がある．

しかし，彼らの手法は，人手評価により，目的/主体役割としての適切さに応じて 0 から 10 までの値でスコア付けした学習データの特徴を有効に活用せず，正例と負例に分け，2 値分類問題として解いている．

そこで，本研究では，この順位尺度を直接利用すべく，ランキング学習を用いた目的/主体役割の自動獲得手法を提案する．また，提案法は，日本語における特質構造を自動獲得する手法であるが，我々の知る限り，日本語における手法は他に存在しない．

2　特質構造

特質構造は，対象概念が持つ性質を以下の 4 つの異なった観点から記述するものである．

構成役割 (constitutive role)：対象を構成する部分や材料，成分

形式役割 (formal role)：より大きな領域の中で対象を区別するのに必要なもの

目的役割 (telic role)：対象の持つ機能や目的

主体役割 (agentive role)：対象の発生や起源に関する事柄

例えば,「本」の特質構造は次のようになる．

$$\begin{bmatrix} \text{本} & & \\ \text{構成役割} & = & \text{表紙, ページ, 紙} \\ \text{形式役割} & = & \text{出版物} \\ \text{目的役割} & = & \text{読む} \\ \text{主体役割} & = & \text{書く, 出版する} \end{bmatrix}$$

これにより,「本」の目的は「読む」ことであり,「書く」ことや「出版する」ことによって作られるといった百科事典的な知識を得ることができる．

構成役割と形式役割に関しては，既存の WordNet などのオントロジーの部分全体関係や上位下位関係を，そのまま用いることができる場合が多く，また，これら部分全体関係や上位下位関係を自動獲得する手法も，すでに数多く研究されている．そこで，本研究では，それ以外の目的役割と主体役割の獲得に焦点を絞る．

3　特質構造の獲得に関する既存研究

特質構造を獲得する既存の研究には，Wenderoth ら [3, 4] と，Yamada ら [5] による手法が存在する．

Wenderoth らは，Web コーパスから，特質構造の各役割ごとに，パターンマッチングを用いて役割になりうる要素を取得し，その要素を対象名詞との共起度順に並べ替えて，役割として適切な順に要素を並べたリストとして獲得する手法を提案している．彼らの手法で高精度を出すためには，パターンマッチングで役割獲得の再現率と精度が高くなるように，何度も試行を繰り返して最適なパターンを探す必要がある．

Yamada らの手法は，機械学習を用いて，コーパスから目的役割と主体役割として適切な順に要素を並べたリストを獲得する手法である．彼らは，まず，30 の名詞に対し，目的役割と主体役割のそれぞれに 50 個の動詞を割り振り，各動詞が名詞に関して役割として適切かどうか，0 から 10 までのスコア付けを人手で行った．次に，スコアに 7 から 10 が付けられた名詞-動詞のペアを正例，スコアが 0 であるペアを負例とし，最大エントロピー法を用いて学習を行い，スコアを出した．素性には，形態素解析および係り受け解析されたコーパス内でその名詞-動詞が共起する文の名詞と係り受け関係にある語の品詞，動詞と係り受け関係のある語の品詞とその係り受け関係を用いている．そして，そのスコア順に要素を並び替え，目的役割，主体役割の順位リストとした．

我々の提案手法は，獲得パターンを人手作成する必要のない Yamada らの手法をベースとしている．しかし，彼らの手法が，0 から 10 までの尺度でスコア付けしたデータの特性を生かさず，2 値分類問題に帰着して解くのに対し，我々の手法は，ランキング学習を用い，データの順位尺度を直接的に利用する．

4 ランキング学習を用いた特質構造の自動獲得

本実験では，人手で作成した学習データと，ランキング学習を用いて，名詞と係り受け関係にある「助詞+動詞」を，目的役割，主体役割として適切な順に並べ替えた順位リストを獲得する．以降では，まず，実験で使用するコーパスと辞書について紹介し，その後，提案手法について述べる．

4.1 実験で使用したコーパスと辞書

4.1.1 日本語係り受けコーパス

日本語係り受けコーパス[1] は，約 1 億件のウェブページを収集した日本語ウェブコーパス 2010 [2]に対して CaboCha (Version: 0.60pre4, 辞書: NAIST-jdic-0.6.3) を用いて係り受け解析をし，助詞を介した語と語の係り受けを抽出したものである．

4.1.2 NTT 日本語語彙大系

NTT 日本語語彙大系は，意味体系，単語体系，構文体系から構成される日本語シソーラスである．意味体系は，「具体」，「人間」，「植物」，「スポーツ」といった意味的な属性を，階層的に分類・体系化したものであり，3,000 個の属性が，最大で 12 段の木構造により体系化されている．単語体系では，30 万語の単語が，意味体系の意味属性により定義されている．

4.1.3 動詞項構造シソーラス

動詞項構造シソーラス (Version: 0.902)[3]は，竹内らによって作成された動詞の概念を整理した辞書である．動詞 4,425 語，7,473 語義について人手で 5 階層の分類がされており，この分類は，細分類で 940 分類となっている．

4.2 学習・評価に用いるデータの作成

まず，学習および評価に用いる名詞として，日本語語彙大系の単語体系に含まれる一般名詞の中から，目的や発生原因が比較的明らかな 90 個の名詞を選択した．選択した名詞の一部は，次の通りである．

> 蛇口, 蛍光灯, トラック, ネクタイ, 封筒, 本, 鞄, 絵, 鉛筆, 家, 機械, 辞書, ジュース, 料理, お茶, クッキー, 財布, 医者, 弁護士, 学生, 大工, スーパー, 病院, 刑務所

次に，係り受け関係にある (名詞, 助詞, 動詞) の三つ組とその出現頻度をまとめた日本語係り受けコーパスのデータから，先ほど選択した名詞のそれぞれについて，名詞と，その名詞と係り受け関係にある「(が | を | に | で) + 動詞」と間の共起度を計算する．共起尺度には，語と語の関連性を計算するのによく使われる Jaccard 係数を用いた．

$$Jac = \frac{|N \cap CV|}{|N| + |CV| - |N \cap CV|}$$

ここで，N は名詞の出現頻度で，CV は「助詞 + 動詞」の出現頻度である．共起度を計算したあと，名詞ごとに，共起度順上位 50 組の「助詞 + 動詞」の組を残す．この時，動詞項構造シソーラスに含まれない動詞を含む組は除く．

4.3 人手による役割らしさスコア付け

名詞と共起度の高い「助詞 + 動詞」の組が得られたのち，90 個すべての名詞について，対応する 50 組の「助詞 + 動詞」のそれぞれが，名詞に対して目的役割，主体役割として適切であるかを，人手で評価し，適切なものほどスコアが高くなるように 0 から 10 までのスコアを付けた．スコアの度数分布を図 1 に示す．

[1] http://hayashibe.jp/jdc/
[2] http://s-yata.jp/corpus/nwc2010/
[3] http://cl.cs.okayama-u.ac.jp/rsc/data/

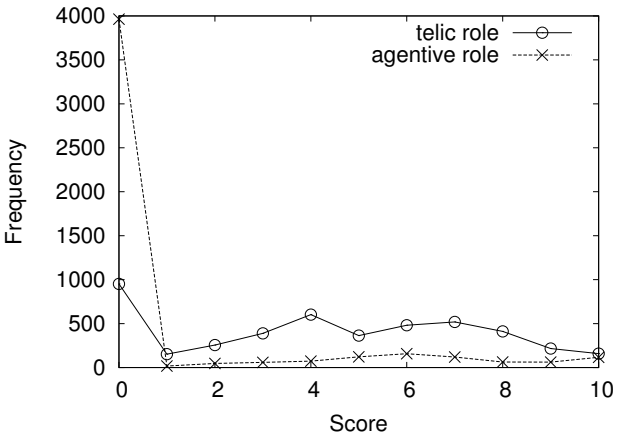

図 1: スコアの度数分布

4.4 目的/主体役割の自動獲得手法

人手で付けた各役割らしさスコアに対して,「共起度で並べ替えた時の順位」,「名詞の意味属性+助詞」,「動詞の分類と動詞の取る深層格の情報+助詞」を素性とし．目的/主体役割らしさのランキング学習を行う．

素性として与える名詞の意味属性は，日本語語彙大系の単語体系で，単語の定義に直接用いられてる意味属性だけでなく，意味体系の木構造上で，その意味属性の上位に位置するすべての祖先ノードの意味属性も含める，例えば,「鉛筆」の素性は,「具体」,「具体物」,「無生物」,「人工物」,「道具」,「文具・おもちゃ等」,「文具」となる．また，複数の意味属性を持つ名詞については，すべての意味属性および意味体系上その上位に位置する意味属性を素性とした．

動詞の素性には，動詞項構造シソーラスから取得した 5 階層の分類と，主格に来る名詞のタイプを用いた．例えば，動詞「買う」であれば，分類「状態変化あり - 位置変化 - 位置変化 (物理)(人物間) - 他者からの所有物の移動-購入」と，主格にくる名詞のタイプ「動作主」を素性とする．複数の語義を持つ動詞については，名詞の場合と同様にすべての語義の分類を素性とした．ランキング学習には，SVMrank[4]を用いた．

5 精度評価および考察

5.1 評価尺度

この研究では，ランク上位の並びに興味がある．そこで，評価には Yamada らの研究 [5] で用いられている順位相関係数を使用した．この順位相関係数は，スピアマン順位相関係数を変形したもので，上位 m 個の順位の相関係数を計算することができる．順位相関係数は次の式により求める．

$$\begin{aligned} Rs &= 1 - \sum_{x=1}^{m} d_x^2 / E\left(\sum_{x=1}^{m} d_x^2\right) \\ &= 1 - 6 \times \sum_{x=1}^{m} d_x^2 / m(2m^2 - 3nm + 2n^2 - 1) \end{aligned}$$

ここで，n はデータの数，m はランク上位の順位相関係数を求めるデータの個数，d_x はデータ集合中の順位の差であり，$E(x)$ は x の期待値を表している．通常のスピアマン順位相関係数と同様に，2 つのデータ集合の順位に完全な正の相関があれば，Rs の値は 1 となり，無相関の場合は 0 となるが，データ集合に負の相関がある場合は，Rs の値が -1 よりも小さくなることもある．

5.2 精度評価および考察

評価対象の名詞が含まれる，(名詞,「助詞 + 動詞」) の係り受け対を残し，その他すべての係り受け対を学習データとして用いる．このようにして，90 個すべての名詞に対して，その名詞と係り受け関係にある「助詞 + 動詞」を並べ替え，目的役割と主体役割の順位リストを獲得する．その後，獲得した順位リストと人手で付けたスコアの並び順との順序相関を上位 1 から 20 個まで，すべての名詞について計算し，平均を求めた．その結果を，目的役割については図 2，主体役割については図 3 に示す．ここで，4.4 節で挙げた，すべての素性を学習に用いた結果を ALL,「動詞の意味分類 + 助詞」を用いなかったものを ALL-VERB,「名詞の意味属性 + 助詞」を用いなかったものを ALL-NOUN, いずれも用いなかったものを ALL-VERB&NOUN とした．ベースラインは，学習データのスコア 0 の係り受け対を負例，スコア 7 から 10 の係り受け対を正例として，LibSVM[5]で学習し，確率推定オプション (-b 1) を指定して，その結果出力された確率が高い順に並べ替えたものとする．素性には，提案法 (ALL) と同じものを用いた．

図 2，3 から，提案法 (ALL) は，上位 N 個 ($1 \leq N \leq 20$) の順位相関係数で，目的役割，主体役割のどちらについても，ベースラインを上回っており，ランキング学習の有効性が示された．上位 3 つの順位相関係数は，目的役割の場合，提案法 (ALL) が 0.789, ベースラインが 0.551 であり，主体役割の場合，提案法 (ALL) が 0.653，ベースラインが 0.516 であった．

[4]SVMrank Version:1.00 (http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html)

[5]LibSVM Version:3.11 (http://www.csie.ntu.edu.tw/~cjlin/libsvm/)

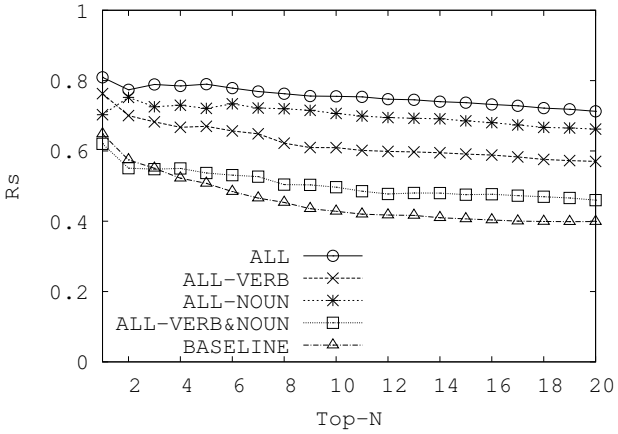

図 2: 獲得した目的役割リストの順位相関

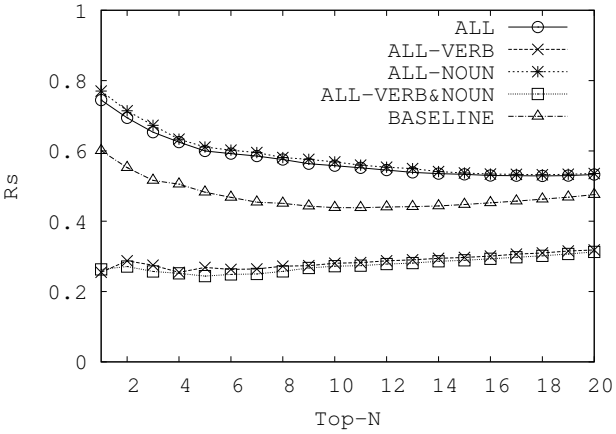

図 3: 獲得した主体役割リストの順位相関

表 1: 獲得した目的役割順位リストの例

順位	辞書		トラック	
1	で 調べる	(6.14)	を 運転する	(5.26)
2	で 引く	(5.30)	で 走る	(4.90)
3	を 使う	(5.02)	で 運ぶ	(4.40)
4	で 確認する	(4.70)	に 乗せる	(3.71)
5	を 利用する	(4.44)	に 積み込む	(3.22)
6	で 検索する	(4.39)	で 輸送する	(3.19)
7	で 確かめる	(4.31)	に 乗る	(3.11)
8	で チェックする	(3.90)	に 詰め込む	(3.10)
9	を 活用する	(3.72)	に 積む	(3.09)
10	を 開く	(3.61)	で 配送する	(3.07)

表 2: 獲得した主体役割順位リストの例

順位	本		お茶	
1	に 書く	(2.74)	を 栽培する	(2.76)
2	を 作る	(2.39)	を 仕入れる	(2.56)
3	を 出版する	(2.02)	を 貰う	(2.34)
4	を 執筆する	(1.93)	を 買う	(2.12)
5	に する	(1.87)	を 摘む	(2.09)
6	を 書く	(1.84)	を 出す	(1.71)
7	に まとめる	(1.50)	を 用意する	(1.64)
8	を 購入する	(1.46)	を 沸かす	(1.63)
9	を 選ぶ	(1.38)	を 準備する	(1.52)
10	が 出る	(1.32)	を 頼む	(1.49)

また，ALL-VERB，ALL-NOUN，ALL-VERB&NOUN の比較から，目的役割の獲得には,「名詞の意味属性」と「動詞の意味分類」のいずれも効果があり，主体役割の獲得には,「動詞の意味分類」は役に立つものの,「名詞の意味属性」は，ほとんど役に立たないことがわかる．

獲得できた目的役割の順位リストの例を表 1 に，主体役割の順位リストの例を表 2 に示す．なお，表の括弧内の数値は，SVMrank の出力スコアである．

獲得できた順位リストと人手でスコア付けした順位の並びで，大きく異なる部分は少なかった．しかし，例えば,「先生」について獲得した目的役割の順位リストにおいて,「が 教える」よりも「が 書く」や「が 作る」などが上位に並ぶなど，対象特有で役割としてふさわしい要素が最上位に来ない場合もあった．この問題は，上位のカテゴリとの間の共起度を計算し，要素が対象特有のものであるかも考慮に入れることによって，解決できるのではないかと考えている．

6 おわりに

本研究では，ランキング学習を用いて，与えられた名詞に対する目的役割と主体役割の順位リストを自動で獲得する手法を提案した．今後の課題としては，項構造や事象構造に関しても，自動獲得もしくは，既存のリソースの変換をし，包括的な生成語彙論的リソースを作成することが考えられる．

参考文献

[1] James Pustejovsky. *The Generative Lexicon*. MIT Press, 1998.

[2] 植村 将人. **生成語彙論に基づく名詞句「*A* の *B*」の意味解釈**. 修士論文, 北陸先端科学技術大学院大学, 2005.

[3] Philipp Cimiano and Johanna Wenderoth. Automatically learning qualia structures from the web. In *Proceedings of the ACL-SIGLEX Workshop on Deep Lexical Acquisition*, pages 28–37, 2005.

[4] Johanna Wenderoth. Automatic acquisition of ranked qualia structures from the web. In *Proceedings of the ACL*, pages 888–895, 2007.

[5] Yamada Ichiro, Baldwin Timothy, Sumiyoshi Hideki, Shibata Masahiro, and Yagi Nobuyuki. Automatic acquisition of qualia structure from corpus data. *IEICE transactions on information and systems*, 90(10):1534–1541, 2007.