名詞句「NP の NP」の意味関係とその統計的性質

田中 省作　飯田 健二　冨浦 洋一　日高 達
(九州大学大学院 システム情報科学研究科)

1 はじめに

日本語文中では, 2つの名詞句が助詞「の」で結合した名詞句「NP の NP」が頻繁に現れる. この名詞句「NP の NP」は, 表層的には単純であるがその意味は多様である. このような名詞句「NP の NP」に対して Montague の形式化に従い統語構造と意味構造を対応させるような文法体系 [4] が提案されている. この文法体系では, 従来, 単一の統語範疇として扱われていた名詞句を意味的観点から4つの統語範疇に細分化し, 名詞句の統語構造と意味構造との関係を対応づけている. この文法体系に基づいて名詞句の意味構造を推定する場合, ある特定の統語構造 (「T の CN」) では, 名詞句同士を意味的に結ぶ関係 (意味関係) を導出する必要がある. 本稿では, 名詞句「NP の NP」で導き出される意味関係を分類し, その統計的性質および統計的決定法について示す.

2 名詞句「NP の NP」

名詞句の統語範疇の細分化

[4] で提案されている名詞句「NP の NP」の文法体系では, 名詞句を4つの統語範疇, 普通名詞句 (CN), 項句 (T), 関係名詞句 (RN), 事象名詞句 (EN) に細分化する. 以下, 簡単にこれらの統語範疇について説明する.

(1) 普通名詞句 (common noun phrase; CN)
性質を表しているような名詞句で, 例えば「色白」「美人」などが CN で, それぞれ "色白である", "美人である" という性質を意味している. また, 統語的には, 「ある」「その」というような限量子が結合することができる.

(2) 項句 (term; T)
ある特定の個体や事象を指示しているような名詞句で, 「太郎」や「その人」というような固有名詞や CN に限量子が付加した名詞句などが T である.

(3) 関係名詞句 (relation noun phrase; RN)
個体や事象間の関係を表す名詞句で, 「兄」や「目的」などが RN である. 例えば, 「兄」という名詞句は, T の「太郎」を「の」で連接して「太郎の兄」という名詞句を構成し, "太郎の兄である人" を指示することになる.

(4) 事象名詞句 (event noun phrase; EN)
動詞性の事象を指示する名詞句で, 「努力」や「考え」というものがある. 動詞が名詞化したような名詞は全てこの範疇に属す.

名詞句の意味構造

名詞句「NP の NP」の意味構造は, 高階論理式で記述され, 統語構造を構成した統語規則に対応した翻訳規則によって形式的に導出される. その意味構造は, 次の3つに大別される.

(1) **統語構造が「CN の CN」の場合**
この場合は, NP_1 の性質と NP_2 の性質を充たすものを指示する. 例えば, 「大型の犬」は, 次のような意味構造になる.

$$\lambda x\ 大型^*(x) \wedge 犬^*(x)$$

(2) **統語構造が「T の CN」の場合**
この場合は, 名詞句の表層表現に明示的には現れていない意味関係 R を推定した上で意味構造は導出される. 例えば, 「駐車場の車」であれば「駐車場に 位置している 車」といえるので, 意味関係として『位置関係』が成り立つと推定される. このとき, 意味構造は,

$$\lambda P \exists y\ 車^*(x) \wedge 駐車場^*(y) \wedge 位置(x,y) \wedge P(x)$$

ただし, '位置 (x,y)' は "x が y の位置にある" ことを表す述語である.

(3) **統語構造が「T の RN」の場合**
この場合は, NP_2 の関係名詞句自身が NP_1 を受けて指示対象を対応づけるような意味関係を内在しており, その関係によって個体が限定される. 例えば,「母」は『親子関係』という意味関係を内在している RN である.「太郎の母」という名詞句の意味構造は,

$$\lambda P\ 母^*(x, 太郎^*) \wedge P(x)$$

と表される. ただし, '母 (x, y)' は "x は y の母親である" ことを表す述語である.

このように統語構造に対応して意味構造が「T の CN」の場合の意味関係 R を除いては決定される. よって, 正しい意味構造を導くには, 名詞の統語構造と適用する生成規則を正しく選択することが大切である.

名詞の統語範疇 T および EN については, 表記や国文法における文法情報によって判別できるが, 残った CN と RN の区別については客観的な基準がなかった. そこで, CN と RN の判別をコーパス中の名詞句の共起情報を基に (半) 自動的に行う手法を既に提案している [3]. [3] では, 名詞句「NP_1 の N_2」の共起関係から名詞 N_2 における名詞句 NP_1 の意味範疇の散らばりを定量化し, CN,RN の (半) 自動的な分類を行っている.

3 意味関係の分類

ここでは, 名詞句の統語構造が「T の CN」の際に推定する意味関係について, 次のように分類する. まず, 多くの名詞句間で成り立つような意味関係,『所有関係』や『位置関係』など 7 つの意味関係を設定し, これらを一般的な意味関係と呼ぶことにする. また,「写真の私」では「写真」と関連性の高い動詞である「写す」を NP_1,NP_2 を意味的に結びつける動詞として「写真に 写っている 私」と考えることができる. このように個々の名詞に固有の動詞によって NP_1,NP_2 を意味的に結びつける意味関係をその他の意味関係と呼ぶことにする.

(1) **所有関係**
NP_1 が NP_2 の所有者であることを表す.
「私の車」,「彼の計算機」

(2) **所属関係**
NP_2 の所属が NP_1 であることを表す.
「九大の職員」,「日本の大使」

(3) **位置関係**
NP_2 のある場所が NP_1 であることを表す.
「机の上のパソコン」,「道端の若木」

(4) **作成-生産関係**
NP_1 の作成者が NP_2 であることを表す.
「彼のプログラム」,「IBM のパソコン」

(5) **数量関係**
NP_1 が NP_2 に関する数量を表している.
「10km の道のり」,「30kg のみかん」

(6) **全体-部分関係**
NP_1 が NP_2 を部分としてもつことを表す.
「木の枝」,「車のタイヤ」

(7) **部分-全体関係**
NP_2 は NP_1 を部分としてもつことを表す.
「赤い屋根の建物」,「髭の男」

(8) **その他の意味関係**
名詞句中に現れる名詞に固有の動詞によって NP_1,NP_2 を意味的に結ぶ関係である.
「写真の人物」⇒『写る』
「彼の理論」⇒『考える』

これらの意味関係が名詞句「NP の NP」においてどの程度の割合を占めているかを, EDR コーパスのテキストに含まれる名詞句「NP_1 の NP_2」(統語構造は「T の CN」) からランダムに 2021 個サンプリングを行い調査した結果, (1)〜(7) の意味関係に属す「T の CN」が 1821 個 (90.1%), 200 個 (9.9%) が (8) のその他の意味関係に属していた.

3.1 一般的な意味関係の推定

一般的な意味関係については, 予め意味関係毎に名詞の共起関係を収集し, 名詞句の意味関係を統計的に推定することが可能と思われる. そこで, 実際に EDR コーパス中で名詞句「n_1 の n_2」が一般的な意味関係 R で出現したという用例 $\langle n_1, n_2, R\rangle$ を基に, ノンパラメトリックな確率密度の推定法で

ある k-NN 推定法を用いた Bayes 判別法による名詞句の一般的な意味関係の推定を行った.

$\mathcal{R}$ を一般的な意味関係の集合とすると, Bayes 判別法では, 名詞句「n_1 の n_2」が発生したときに, n_1, n_2 を意味的に結ぶ一般的な意味関係が $R \in \mathcal{R}$ となる確率 $P_r(R|\lceil n_1 \text{ の } n_2 \rfloor)$ を計算し,

$$\underset{R \in \mathcal{R}}{argmax} P_r(R|\lceil n_1 \text{ の } n_2 \rfloor)$$

を名詞句「n_1 の n_2」の意味関係として決定する. さらに, Bayes の定理より,

$$\underset{R \in \mathcal{R}}{argmax} P_r(R) P_r(\lceil n_1 \text{ の } n_2 \rfloor|R)$$

となる. ここで, $P_r(R)$ は,

$$P_r(R) = \frac{(\text{意味関係が } R \text{ の用例数})}{(\text{全用例数})}$$

と推定する. また, $P_r(\lceil n_1 \text{ の } n_2 \rfloor|R)$ は k-NN 推定法 [2] で推定する.

Definition (k-NN 推定法)
大きさ N のサンプル $\mathcal{S}^N$ における $\boldsymbol{x}$ の確率密度の推定値は,

$$\hat{p}(\boldsymbol{x}; \mathcal{S}^N) = \frac{k-1}{N} \frac{1}{A(k, \mathcal{S}^N, \boldsymbol{x})}$$

である. ただし, $A(k, \mathcal{S}^N, \boldsymbol{x})$ は $\boldsymbol{x}$ と $\mathcal{S}^N$ における $\boldsymbol{x}$ の k-nearest neighbor との距離 $r(k, \mathcal{S}^N, \boldsymbol{x})$ を半径とする超球の体積である. ◇

ここで入力を名詞句「NP_1 の NP_2」のそれぞれの主辞 $\langle n_1, n_2 \rangle$ とし, 単語間の距離を,

$$\Delta(\langle n_1, n_2 \rangle, \langle n_1', n_2' \rangle) = \sum_{i=1}^{2} \delta(n_i, n_i')$$

とする. また, $\delta(n, n')$ は EDR シソーラス上の名詞 n, n' の概念間の最短パス長である.

その結果, 1821 個の $\langle n_1, n_2, R \rangle$ の用例をサンプルとして分類実験を行った結果, 全意味関係の平均の正解率は 78%, 第 3 位までの累積認識率では 93%となった.

3.2 その他の意味関係の統計的性質

その他の意味関係となりえる動詞は, 名詞句中に現れる名詞そのものに非常に強く依存する. 例えば,「写真の私」を結びつける動詞は『写っている』だが,「写真」と同じ創作物であっても「絵の私」では『描かれている』となる. このように動詞性の意味関係は,『具体物』や『抽象的なもの』といった高位のレベルの概念からは推定されにくく, 各単語レベルで意味関係になりえる動詞の候補を抽出する必要がある.

そこで, 予備的に次のような事柄について調査を行った.

特定の動詞との共起関係の強さ

名詞句「NP_1 の NP_2」では, その意味関係は明示的に表層表現には現れないが, 人はほぼ正しく一意にその関係を決定できることから, 意味関係になる動詞が NP_1 または NP_2 と強く共起しているのではないか, ということが予想される. また,「炎の絵」と言った名詞句を考えた場合, その意味を結ぶ『描かれた』という動詞は NP_2 の「絵」とは強く共起するが,「炎」とはほとんど共起しない. よって, 意味関係となる動詞 v は, NP_1 または NP_2 のいづれか一方から共起関係の強い動詞が引き出されていると考えられる. そこで, 単文中の名詞 n および動詞 v の共起関係 $\langle n, v \rangle$ から, 名詞 n に共起関係の強い特定の動詞が存在する度合をエントロピーを用いて定量化する. 共起関係 $\langle n, v \rangle$ から名詞 n に対する各動詞の分布 $P_r(v|n)$ を用いて, 特定の動詞との共起関係の強さ $\mathcal{H}_V(n)$ を次のように定義する. ただし, v は動詞で V は動詞の集合である[1].

$$\mathcal{H}_V(n) = -\sum_{v \in V} P_r(v|n) \log_{|V|} P_r(v|n)$$

ただし,

$$P_r(v|n) = \frac{f(\langle n, v \rangle)}{\sum_{v \in V} f(\langle n, v \rangle)}$$

エントロピーは, そもそも不確実性を表すものであるから, 名詞 n についてある特定の動詞との共

[1] 通常, エントロピーは log の底として 2 を割り当てるが, log 関数は底が 1 より大きければ単調増加関数であるので, $\mathcal{H}_V(n)$ の最大値が 1 になるように V の要素数 $|V|$ を底とした. $\mathcal{H}_V(n) = 1$ となるのは, V 中の全ての動詞 v で $P_r(v|n) = 1/|V|$ のとき.

起関係があれば0に近い値を, 多くの動詞と共起関係がある場合は1に近い値をとる. よって, 名詞句「NP_1 の NP_2」のその他の意味関係が, 名詞との共起関係が高いという仮定から, $\mathcal{H}_V(n_i)$ のより小さな値をとる名詞から意味関係の動詞が導き出されやすいのではないかと考える.

実際に, EDR コーパス中の名詞句「NP_1 の NP_2」で予め人手で動詞性の意味関係が付与されている名詞句 200 個について, NP_1,NP_2 のいづれから意味関係の動詞が導き出されているかと, $\mathcal{H}_V(n)$ との関連性について調べた. その結果, 83%にあたる 166 個の名詞句については, $\mathcal{H}_V(n)$ が低い名詞から, 意味関係になった動詞と強い共起関係が確認された.

この傾向は, $\mathcal{H}_V(n)$ を動詞性の意味関係の推定の際の候補の絞り込みなどに援用できるのではないかと考えられる.

格関係の考慮

名詞句「NP_1 の NP_2」は, 「NP_1 の」を連体修飾句と置き換えて,

$$「NP_1\ c_1\ v\ NP_2\ (c_2)」$$

と言い換えることができる. 例えば, 「私の電車」では「私 が 乗車した 電車 (に)」と言い換えができる. つまり, 深層的に名詞「私」「電車」が, それぞれガ格, ニ格で動詞「乗車する」と係り受け関係を結んでいる. ただし, (c_2) は格 c_2 が表層的には現れていないことを表す. このように深層的に名詞 n_1, n_2 がそれぞれ c_1, c_2 格で, ある動詞 v と係り受け関係を結んでいることを $\langle\langle n_1, n_2, c_1, c_2, v\rangle\rangle$ と書くことにする. そこで, 名詞 n_1, n_2 が「n_1 の n_2」という形で出現したとき, 動詞 v とそれぞれ c_1, c_2 格で係り受け関係を構成する確率,

$$P_r(\langle\langle n_1, n_2, c_1, c_2, v\rangle\rangle | 「n_1 \text{ の } n_2」)$$

を計算することにより, 名詞句「NP_1 の NP_2」における意味関係となりえる動詞 v および名詞が埋める格 c_1, c_2 の候補を絞りこめることができるのではないかと思われる. しかし, 現時点では $\langle\langle n_1, n_2, c_1, c_2, v\rangle\rangle$ について十分な用例を用意できない. そこで, 現在, 単文中の名詞-格-動詞の共起関係で近似した小規模な実験を行い, 比較的良好な結果を得ている [1].

4 おわりに

本稿では, 名詞句「NP の NP」における意味関係を分類し, 多くの名詞句で成り立つ一般的な意味関係については, k-NN 推定法を用いて意味関係の推定が比較的高い正解率で分類できることを確認した. また, 名詞固有の動詞で意味関係を結ぶ動詞性の意味関係について, 幾つかの統計的性質について調査した. 今後は, エントロピー $\mathcal{H}_V(n)$ と格関係の分布を考慮した意味関係の候補の動詞の絞り込みについて実験を行いたい.

また, 本稿では一般的な意味関係と動詞性の意味関係を区別して扱ったが, 統計的性質を確率係り受け文脈自由文法などに組み込み, 統一的に名詞句の意味関係を決定する機構を構築していきたい.

参考文献

[1] 飯田 健二; 名詞句「NP の NP」の意味解析 – 統計的手法を用いた意味関係の抽出–, 九州大学 修士論文, 1998

[2] Keinosuke Fukunaga; Introduction to Statistical Pattern Recognition, ACADEMIC PRESS INC., 1972

[3] 田中 省作, 冨浦 洋一, 日高 達; 名詞句「NP の NP」の意味構造推定のための知識獲得 – 統計情報を用いた普通名詞/関係名詞の分類–, NLP シンポジウム「大規模資源と自然言語処理」,1997

[4] 冨浦 洋一, 中村 貞吾, 日高 達; 名詞句「NP の NP」の意味構造, 情報処理学会論文誌, Vol.36, No.6, 1995