社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (15)
2008－ＮＬ－183 (15)
2008／1／22

枝分かれ同時確率モデルを用いた「A の B」の意味分類

森山 健太† 但馬 康宏† 藤本 浩司‡ 小谷 善行†

†東京農工大学 工学府 情報工学専攻

‡テンソル・コンサルティング株式会社

概要

日本語における多義性を持つ表現として，「A の B」という表現がある。「A の B」とは，連体助詞の「の」が名詞を伴って「～の」という文節を構成し，その文節が体言に係る場合を指す。「A の B」は文中における単語 A と単語 B の関係によって様々な意味になりうる。「A の B」は，文中に高い確率で出現するため，意味解析の上で重要な問題である。現在，専用の辞書を用いる方法や国語辞典を用いる方法等，様々な方法が提案されている。

本研究では，木構造の概念知識を用いた機械学習を用いて，「A の B」の意味を推定することを目的としている。単語の概念共起確率を上位概念の共起確率の積によって近似する，枝分かれ同時確率モデルを用いて機械学習を行った結果，19 択問題において 67.2%の精度を得た。

Classification of "A no B" using Nested Joint Probability Model

Kenta MORIYAMA†, Yasuhiro TAJIMA†, Koji FUJIMOTO‡ and Yoshiyuki KOTANI†

†Tokyo University of Agriculture and Technology

‡Tensor Consulting Co. Ltd

Abstract

Noun phrase "A NO B" is one of an expression with polysemy in Japanese. "A NO B" contains of adnominal particle "NO" that composes the clause with noun A, and it modify noun B. The meaning of "A NO B" is influenced by the relation between A and B. "A NO B" is a big problem on the semantic analysis because it appears in the sentence at a high probability. Various methods to solve it have been proposed up to now. For example, there are a method using a special dictionary, and other method using the national language dictionary.

We use tree structure Thesaurus to analysis "A NO B" with machine learning. The machine learning method is Nested Joint Probability Model that can calculate co-occurrence probability of A and B from their superordinate conceptions. As a result, we obtained 67.2 percent precision.

1．はじめに

本研究では，自動学習により，文中の連体助詞「の」の意味を推定することを目的とする。連体助詞「の」とは，名詞を伴って「～の」という文節を構成し，その文節が体言に係る場合（以後「A の B」と呼ぶ）の「の」の事である。「A の B」は，文中の単語 A と B の関係によって，様々な意味になりうる。

「Ａの Ｂ」は，日本語の文中に非常に高い確率で出現するため，意味解析の上で重要な問題であ

り，現在様々な手法が提案されている。
語彙項目を名詞毎に与え，ＡとＢの語彙項目の組み合わせにより意味を解析する[1]や，国語辞典を用いた解析を行う[2]のような研究がある。

本研究は，事前に意味分類のルールを定めずに，EDR 概念辞書[3]から得た前後の単語の概念知識を用いた機械学習によりルールを生成するシステムを作成することを目的としている(図1)。

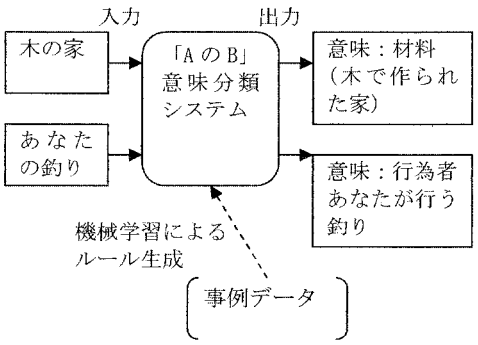

図１　機械学習によるＡのＢ意味分類システム

[4]では同様の概念知識を用いた決定木学習により 61.5%の精度を得たが，本研究では藤本らによって提案された枝分かれ同時モデル[5]を用いて，さらなる精度の向上を目指す。

２．「ＡのＢ」の意味定義

国語辞典を基に「ＡのＢ」の意味を 19 種類定義した。定義の一覧を表１に示す。

表１　「ＡのＢ」定義一覧

意味名	「A の B」における意味	例
所有	A が持っている B	あなたのカバン
所属	A に所属する B	自民党の人
所在	A にいる（ある）B	神奈川の彼
行為の場所	場所 A における行為 B	インドの生活
時	時 A における B	春のイベント
作者	A が作った B	私の論文
行為者	A がした B	彼の釣り
関係・資格	～と A の関係にある B	左のもの
同格	A である B	政治家の太郎
性質・状態	性質 A である B	軟体の動物
材料	A でできた B	鉄の棒
数量・順序・割合	A 個（番目）の B	３個の石
対象	A に対する行為 B	害虫の駆除
所有属性	A がもつ性質 B	私の年齢
A の関係	A と B という関係にあるもの	それの左
分野・内容	内容，分野が A である B	サンマの煙
部分	A が含む部分 B	刀の刃
A の状態	A が状態 B である	あなたの危険
分類	A に分類される B	類人猿の動物

「A の B」をこれらの意味に分類することを目的とする。

３．概念知識

本章では，学習要素である概念知識について説明する。

3.1 概念知識の構造

概念知識とは，EDR 概念辞書に記載されている概念のことを指す。この概念知識は，図 1 のように木構造で記述されている。

各概念にはその概念に最も意味が近い単語が，概念見出しとして記述されている。本研究では，単語と概念見出しが一致した概念を，単語の概念とする。下位概念はより細かい概念を，上位概念はまとまった概念を表し，最も上位の概念は「全ての概念」である。

本研究では「全ての概念」を 0 層概念，その下位概念を 1 層概念とする。また，概念辞書に載っていない単語の概念は「null」とした。

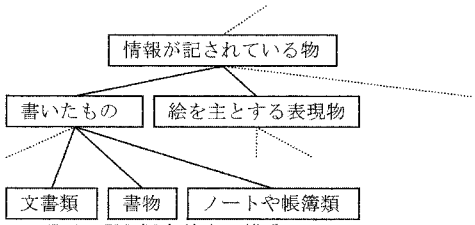

図 1　EDR 概念辞書の構造

3.2 概念知識の学習

単語は多義性をもつため，1 つの単語に対し複数の概念が存在する。例として「犬」という単語には，「犬という動物」「警察などの回し者」という概念が存在する。「A の B」=「犬の餌」の場合の学習要素を表 2 に示す。

表 2　「犬の餌」の学習要素

	A：犬の概念		B:餌の概念
1 層	ものごと	ものごと	ものごと
2 層	もの	もの	もの
3 層	具体物	具体物	具体物
4 層	生命体	生命体	静物
5 層	人間	動物	機能で捉えた具体物
6 層	状態や評価で捉えた人間	種で捉えた動物	飲食物
7 層	立場の評価で捉えた人間	脊椎動物	えさ
8 層	警察などの回し者	哺乳類	動物をおびき寄せて捕えるための食べ物
9 層		犬という動物	

以後，単語 A の n 番目の概念を G_A^n と表記する。単語「犬」の例では，$G_{犬}^1$=「警察等の回し者」，$G_{犬}^2$=「犬という動物」となる。

このような木構造概念において，単語自体の概念(最下位概念)は非常に細かい分類であり，値がスパースとなってしまう。そこで，その上位概念を用いて学習を行う方法をとった。

しかし，どの層の概念を用いれば良いかは事例によって異なる。例として「犬の重量」について，犬の 2 層概念である「もの」と，重量の 5 層概念である「具体物の属性」を見れば意味が所有属性であることが推測できるが，「犬の首輪」の意味が分野・内容であることを推定するには，犬の 5 層概念である「動物」と首輪の 5 層概念の「動物の首にはめる輪」が必要になる。

本研究では，概念の共起確率を枝分かれ同時モデルを用いて適切な上位概念の組によって近似することによって機械学習を行う。

4．概念共起確率による意味分類

「A の B」の意味が M であり，単語 A と単語 B が持つ複数の概念の中で，文中で意図する概念がそれぞれ G_A^n と G_B^m である確率は，ベイズの定理により式(1)のように表すことができる。

$$P(M,G_A^n,G_B^m \mid A\mathcal{O}B) = \frac{P(A\mathcal{O}B \mid M,G_A^n,G_B^m)P(M,G_A^n,G_B^m)}{P(A\mathcal{O}B)} \quad (1)$$

さらにベイズ推定において分母が省略できる。そして，概念が定まれば、概念見出しである単語が定まるため、$P(A\mathcal{O}B \mid M,G_A^n,G_B^m)=1$ となる。よって、式(2)のように変換できる。

$$\arg\max_{M,n,m} P(A\mathcal{O}B \mid M,G_A^n,G_B^m)P(M,G_A^n,G_B^m) = \arg\max_{M,n,m} P(G_A^n,G_B^m \mid M)P(M) \quad (2)$$

後述する枝分かれ同時確率モデルにより，概念共起確率は上位概念の組の積である分解表現式で近似される。近似後の共起確率を Q とすると，最終的な推定式は式(3)となる。

$$\arg\max_{M,n,m} Q(G_A^n,G_B^m \mid M)P(M) \quad (3)$$

また，文献[4]の決定木により，概念知識を学習させる際に，Chasen[6]による形態素解析結果である A と B それぞれの品詞情報も重要な学習要素であると分かっている。品詞情報は表 3 に示したような種類があり，概念が null である事例等の解析に有効である。

表 3 ChaSen による名詞分類例(IPA 品詞体系)

名詞-サ変接続
名詞-ナイ形容詞語幹
名詞-一般
名詞-形容動詞語幹
名詞-固有名詞-一般
名詞-固有名詞-人名-姓
名詞-固有名詞-地域-国
名詞-数
名詞-接尾-人名
名詞-接尾-地域
名詞-代名詞-一般
名詞-副詞可能

そこで，式(3)に品詞情報を単純な重みとして加えた推定式(4)を定義し，この式による意味推定の実験も行った。

$$\underset{M,n,m}{\operatorname{argmax}} P(H_A, H_B \mid M) Q(G_A^n, G_B^m \mid M) P(M) \quad (4)$$

5．枝分かれ同時確率モデル

本章では，枝分かれ同時確率モデルを用いた概念共起確率の上位概念による近似について説明する。

5.1 分解表現式

枝分かれ同時確率モデルは，藤本らによって提案された，枝分かれ構造を持つ 2 事象に対し，その同時確率をそれぞれの上位層を用いた分解表現式で表すことができる確率モデルである。これを概念共起確率に適用する。

2 単語の概念を 1 つずつ選んだものを A,B とする。A の i 層概念を A_i，B の j 層概念を B_j とする。水準 A_i ,B_j の交互作用 I_{AiBj} を以下のように定義する。

$$I_{AiBj} = \frac{P(A_i B_j \mid A_{i-1} B_{j-1})}{P(A_i \mid A_{i-1} B_{j-1}) P(B_j \mid A_{i-1} B_{j-1})} \quad (5)$$

交互作用 $I_{Ai,Bj}$ は，概念 A_i と B_j の上位概念である，A_{i-1} と B_{j-1} が共起した条件のもとでの A_i，B_j が共起する確率の独立性を示す。A_{i-1} と B_{j-1} が共起したという前提において A_i，B_j が独立に生起するならば I_{AiBj} は値 1 をとり，共起しやすければ大きく，共起しにくければ小さくなる。

概念 A と概念 B の最下位概念(単語自体の概念)の層数をそれぞれ m,n とする。A_m，B_n の共起確率 P(AmBn)は概念の包含関係により，

$$P(A_m B_n) = \frac{P(A_{m-1} B_n) P(A_m B_{n-1})}{P(A_{m-1} B_{n-1})} \mathrm{I}_{\mathrm{AmBn}} \quad (6)$$

と表される。この式において $I_{Ai,Bj}=1$ ならば，

$$P(A_m B_n) = \frac{P(A_{m-1} B_n) P(A_m B_{n-1})}{P(A_{m-1} B_{n-1})} : if \quad I_{AmBn} = 1 \quad (7)$$

が成り立つ。

さらに右辺に表われる確率に，順次式(6)を適用することにより，

$$P(A_m B_n) = \frac{P(A_m B_{n-1}) P(A_{m-1} B_n)}{P(A_{m-1} B_{n-1})} : if \quad I_{AmBn} = 1$$

$$= \frac{P(A_m B_{n-2}) P(A_{m-1} B_{n-1}) (A_{m-2} B_n)}{P(A_{m-1} B_{n-2}) P(A_{m-2} B_{n-1})}$$

$$: if \quad I_{Am-1Bn} = 1, I_{AmBn-1} = 1, I_{AmBn} = 1$$

$$= \frac{P(A_m B_{n-2}) (A_{m-2} B_n)}{P(A_{m-2} B_{n-2})} \quad (8)$$

$$: if \quad I_{Am-1Bn-1} = 1, I_{Am-1Bn} = 1, I_{AmBn-1} = 1, I_{AmBn} = 1$$

.....

と展開していくことができる。このようにして得られる上位概念の共起確率の積で表現される式を，P(AmBn)の分解表現式と呼ぶ。これを束（ラティス）で表したものを図 2 に示す。

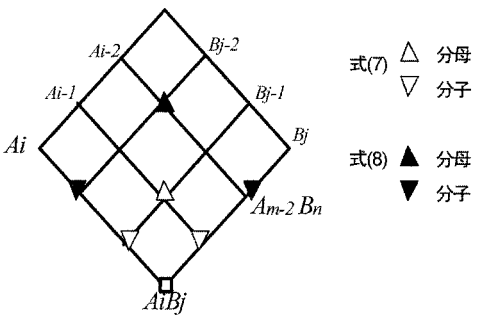

図 2 分解表現式を表すラティス

このように再帰的に式を適用していくことにより，概念の共起確率を交互作用が 1 とみなせる範囲内の上位概念の共起確率の組で表すことができる。

5.2 最適な分解表現式の選択

1 と見なせる上位概念の交互作用が決まれば，

最適な分解表現式を選択することができる。本研究では，交互作用が 1 と見なせるか χ^2 検定によって決定した。学習データから概念 A_i と B_j が共起する事例の観測度数 F_{ij} が与えられているとする。I_{AiBj}=1 なる帰無仮説に対する χ^2 値は，

$$\chi^2 = \frac{F_{i-1j-1}(F_{ij}F_{\bar{i}\bar{j}} - F_{i\bar{j}}F_{\bar{i}j})^2}{F_{i-1\bar{j}}F_{i-1\bar{j}}F_{\bar{i}j-1}F_{\bar{i}j-1}}. \quad (9)$$

where
$$F_{i\bar{j}} = F_{ij-1} - F_{ij}.$$
$$F_{\bar{i}j} = F_{i-1j} - F_{ij}.$$
$$F_{\bar{i}\bar{j}} = F_{i-1j-1} - F_{ij-1} - F_{i-1j} + F_{ij}.$$

と表され，これは自由度 1 の χ^2 分布となる。本研究では，危険率 0.1%の χ^2 検定を採用した。

実際に最適な分解表現式を選択する手順を図 3 に示す。

step1. 候補リスト，確定リストを空にする。
step2. 候補リストに最下位概念の組(Am, Bn)を与える。
step3. 候補リストから一つの組を抜き出し，帰無仮説「交互作用=1」について検定する。帰無仮説が棄却された場合，その組を確定リストに加える。棄却されなかった場合は，上位概念の組に置き換え，それらを候補リストに加える。
step4. 候補リストが空になるまで，step3 から繰り返す。

図 3　最適な分解表現式決定アルゴリズム

この手順により，交互作用=1 とみなせる範囲内で再帰的に上位概念の組に置き換えられ，最適な分解表現式を構成する概念の組が確定リストとして得られる。

5　評価実験

5.1　実験方法

Web から集めたテキストデータに ChaSen を用いて形態素解析し，1800 件の「A の B」を抽出した。「A の B」の意味は，人手で値を付けたものを正解とした。データを 5 分割し，クロスバリデーションで実験を行った。

文中で意図されている概念だけを人手で選択し，学習させることが好ましいが，本研究ではその作業を省き単語が持つ全ての概念を学習データとした。

そのような学習データにおいて，m 層の概念 Am と n 層の概念 Bn が共起する確率 P(Am, Bn)を以下の式によって近似し，計算した。式中の AllCount(S)は，学習事例集合 S に含まれる事例数を表す。F(m, Am, n, Bn, S)は，学習事例集合 S の中で，A の概念集合に m 層が Am である概念が含まれており，かつ B の概念集合にも n 層が Bn である概念を含む件数を表す。式中の X_m^s は事例 s における A の m 層概念集合であり，Y_n^s は事例 s における B の n 層概念集合を表す(図 4)。

$$P(A_m, B_n) \approx \frac{F(m, A_m, n, B_n, S)}{AllCount(S)}. \quad (10)$$

$$F(m, A_m, n, B_n, S) \equiv |\{s \in S \mid A_m \in X_m^s, B_n \in Y_n^s\}| \quad (11)$$

事例 s				事例…
	A：犬の概念		B: 餌の概念	A：…
1 層	ものごと	ものごと	ものごと	…
2 層	もの	もの	もの	…
3 層	具体物	具体物	具体物	…
4 層	生命体	生命体	静物	…
5 層	人間	動物	機能で捉えた…	…
6 層	状態や評価で捉…	種で捉えた動物	飲食物	…
…	…	…	…	…

X_5^s　　Y_4^s

図 4　概念集合 X_m^s, Y_n^s の例(X_5^s, Y_4^s)

5.2 実験結果

品詞共起確率によって重み付けを行うもの（式 4）と，行わないもの（式 3）を比較した。また，比較対象として文献[4]の概念曖昧性に対応した決定木学習システムを用意し，比較した。決定木学習システムの学習要素は，A と B それぞれの概念知識(1～10 層)，品詞，表層語，読みである。これらのシステムに対し，テストデータ全てにおける精度と，概念辞書に単語が載っていた事例を対象とした場合の精度を測定した。ベースラインは最も多い意味である「所有属性」が占める割合 24.9%である。実験結果を表 4 に示す。

表 4　実験結果の精度(%)

	全て	null を除く
決定木システム	63.0	67.3
枝分かれ同時確率	57.8	60.9
枝分かれ同時確率+品詞	67.2	69.4

5.3 考察

枝分かれ同時確率を品詞共起確率と組み合わせることにより，決定木システムを 4%上回る結果が得られたことから，枝分かれ同時確率モデルによる概念知識の学習が，単語間の関係の推定に対して有効であるといえる。品詞情報を加えない場合の枝分かれ同時確率モデルの精度が低いのは，決定木システムが概念知識，品詞，表層語，読みを学習するのに対し，枝分かれ同時確率モデルの学習要素は概念知識だけであることが原因と考えられる。

また，単語が持つ概念を全て学習データとして与えたため，多量のノイズを含むデータとなってしまっている。学習データの単語概念の中で，文中で意図されていない概念を削除すれば精度は向上すると考えられる。

本研究では品詞情報を加える手法として共起確率を単純に重みとして加えただけであるが，この式は A と B の品詞共起と概念の共起が独立でなければ，正確な確率式の近似とはいえない。よって精度の向上のためには，品詞のような概念以外の単語情報を加えた場合に，より正確な近似式となるモデルを実装する必要がある。

6．おわりに

本研究により，枝分かれ同時確率モデルを用いた概念知識の学習は単語間の関係を学習するのに有効であることがわかった。今後の課題としては，概念知識以外の要素を確率モデルに加える事が挙げられる。また，「A の B」以外にも，複合名詞の意味のような，単語間の関係によって答えが定まる問題に対して本システムは有効であると考えられる。

参考文献

[1]植村将人：生成語彙論に基づく名詞句「の」の意味解釈，北陸最先端科学技術大学院大学修士論文，(2005)

[2]黒橋禎夫，酒井康行：国語辞典を用いた名詞句「A の B」の意味解析 情報処理学会研究会 自然言語処理 129-16，pp.109-116 (1999)

[3]日本電子化辞書研究所：EDR 電子化辞書使用説明書(1995)

[4] 森山健太，古宮嘉那子，但馬康宏，小谷善行：概念知識を用いた連体助詞「の」の決定木による自動意味分類，情報処理学会第 69 回全国大会講演論文集，6Q-2(2007)

[5] 藤本浩司，乾伸雄，小谷善行：枝分かれ構造をもつ同時確率モデルによる形態素解析，情報処理学会論文誌 vol.39，No.7pp.2101-2111(1998)

[6] 奈良先端科学技術大学院大学自然言語処理学講座：日本語形態素解析器ChaSen
http://chasen.aist-nara.ac.jp/hiki/ChaSen/