ナイーブベイズ法を用いた意味役割付与に関する実験的考察

岩澤　拓未†　　杉本　徹‡

芝浦工業大学大学院 理工学研究科†

芝浦工業大学　工学部‡

１．研究背景と目的

意味解析は機械翻訳や情報検索などに利用されている．したがって，これらの応用処理を円滑に行うためには，意味解析の精度を高めることが重要となる．意味解析の処理は語義曖昧性解消と意味役割付与の二つに分けられる．本研究ではこれらの処理のうち，概念間に深層格を付与する意味役割付与を扱う．

自然言語の意味解析ライブラリとして，本研究室で開発中のSEM[1]がある．SEMでは意味役割付与にEDR電子化辞書[2]に含まれる日本語共起辞書，および日本語動詞共起パターン副辞書を利用している．これらの辞書には，係り受け関係になりうる概念とその概念間の深層格が記述されている．EDR電子化辞書には，この他に日本語コーパスがある．これは大量の実際の日本語文に対して概念や深層格などの意味的情報が付加されたものである．したがって，より網羅的な解析を行うためには，解析に日本語コーパスを用いることも有効であると考えられる．

本研究では，コーパスを用いた統計的な機械学習による意味役割付与を行う．これにより，高精度で汎用的な意味解析を実現することを目的とする．

２．研究の概要

本研究では，日本語コーパスを基に以下の手順で研究を進めた．

① 日本語コーパスから係り受けデータを抽出する．
② 学習用データを用いて機械学習を行う．
③ 調整用データを用いてパラメータの値を最適化する．
④ テストデータを用いて意味役割付与の評価実験を行い，提案手法の有用性を検証する．

３．研究結果

３．１．係り受けデータの抽出

EDR 電子化辞書の日本語コーパスから文の係り受けデータを抽出した．その結果，207,802 文のデータが得られ，そのうち正常に抽出できたのは207,080 文であった．これを分割し，テストデータとして 20,000 文，調整用データとして 20,000 文，学習データとして 167,080 文を使用した．

３．２．機械学習

機械学習の手法にはナイーブベイズ法を用いた．これは設計が単純であるにもかかわらず，様々な問題に対して，ある程度の精度を持つためである．また，独立性の仮定を行うことによって，学習用データが少なくて済むという利点もある．ナイーブベイズ法を用いた確率値の計算式を以下に示す．

$$P(d \mid F) \tag{1}$$

$$= \frac{P(d)P(F \mid d)}{P(F)} \tag{2}$$

$$= \frac{P(d)\prod_{f_i \in F} P(f_i \mid d)}{P(F)} \tag{3}$$

式(1)において，d は求める深層格を表す．また，F は入力文から得られる素性の集合であり，係り元文節の主辞の単語や表層格などが要素となる．つまり，式(1)は素性集合 F から深層格 d を予測するモデルである．式(2)は式(1)をベイズの定理で変形したものである．次に，$P(F \mid d)$において，F 中の各素性 f_i の出現は互いに独立であると仮定すると，式(3)のように変形される．本研究では，式(3)の確率を最大にするような深層格 d を選択することで意味役割付与を行った．この際に，$P(F)$の値は深層格 d によらず一定であるので省略した．よって，

$$\arg\max_{d \in D} P(d)\prod_{f_i \in F} P(f_i \mid d) \tag{4}$$

を計算すればよい．

式(4)から，推定するべき確率モデルは $P(d)$と $P(f_i \mid d)$である．$P(d)$の推定には，最尤推定法を用いた．また，ディスカウンティングには予期尤度推定法を用いた．ディスカウンティングとは，ある事象の出現回数の値を補正することである．$P(f_i \mid d)$は，

$$P(f_i \mid d) = \beta P_{ML}(f_i \mid d) + (1-\beta)P(f_i) \qquad (5)$$

のように，まず線形補完法を用いて $P_{ML}(f_i \mid d)$と $P(f_i)$の線形和で表すことにより，ディスカウンティングを行った． $P_{ML}(f_i \mid d)$と $P(f_i)$は，共に最尤推定法で推定した．$P(f_i)$については，さらに加算スムージングを用いてディスカウンティングを行った．

上記の方法で，学習データを用いて確率値を計算した．使用する素性は，係り側と受け側の概念や単語の字面，品詞，表層格の組合せを数パターン試した結果，以下の３つの組合せを使用することにした．

- 係り側の概念
- 受け側の概念
- 表層格

３．３．パラメータ β の最適化

一般に統計的な確率値を扱う際には，学習データでの出現頻度が極端に低いと信頼性のある確率値を得られないという問題がある．このような問題はゼロ頻度問題やスパースネスの問題と呼ばれ，ディスカウンティングという出現頻度の補正処理を行う必要がある．提案手法でもディスカウンティングをいくつか行っている．このうち式(5)で表される線形補完法では，二つの確率の重みとしてパラメータ β が与えられている．この値は，実際に調整用データを用いて β の値を変動させ，精度が最も高くなるときの値を採用することで最適化した． その結果，精度が最も高くなったときの β の値は 0.7 であった．精度の計算には以下の式を用いた．

$$\text{精度} = \frac{\text{適切な深層格を付与した係り受け数}}{\text{総係り受け数}} \qquad (6)$$

３．４．評価実験

提案手法が有効であるかを検証するために，テストデータを用いて意味役割付与の実験を行った．ベースラインとしては，学習用データ中に最も多く出現した深層格である“object”を常に付与する手法を用いた．

表 1 に実験結果を示す．表 1 から，提案手法はベースラインよりも高い精度であることがわかった．また，精度の計算には式(6)を用いており，総係り受け数は 162,855 個である．

表 2. 実験結果

	精度（%）
提案手法	66.65
ベースライン	25.62

次に，係り側と受け側の概念の組合せ別の誤り率を出現回数の多かったものから順に表 2 に示す．出現回数はテストデータに出現した回数であり，そのうち深層格の付与結果に誤りがあったものの割合が誤り率である．また，概念の欄では，EDR 電子化辞書における概念 ID の右に括弧付きで概念の説明を付した．表 2 から，「ある状態になる」と「ある状態になる」の組合せのときに出現回数・誤り率ともに大きくなっていることがわかった．

表層格別の誤り率を，誤り率が高いものから順に表 3 に示す．表 3 から，表層格「して」や「で」において出現回数・誤り率がともに大きくなっていることがわかった．

表 1. 概念の組合せ別の誤り率（抜粋）

係り側概念	受け側概念	出現回数	誤り率（%）
102（2）	0e24a0（年という，ものの古さを表す単位）	52	5.8
3ceae3（ある状態になる）	3ceae3（ある状態になる）	50	74.0
101（1）	0e24a0（年という，ものの古さを表す単位）	47	6.4
3cf180（物事の状態）	3ceae3（ある状態になる）	44	20.4
0f6180（昭和という日本の元号）	3c2be7（365 日か 366 日から成り，12 カ月に区分されている，グレゴリオ暦による一年）	43	16.3
0e24a0（年という，ものの古さを表す単位）	3cf6d9（基準になる時刻や時期より前であること）	40	47.5
3ce5fc（はっきりしていて，確かであるさま）	3d06c7（ある状態にする）	38	21.1

表 3. 表層格別の誤り率（抜粋）

表層格	出現回数	誤り率（%）
ら	318	75.2
か	285	74.7
して	1,009	74.1
での	249	55.0
にも	428	54.4
した	742	53.4
で	4,537	53.3

４．考察

４．１．実験結果の考察

表 1 より，提案手法では約 67%の解析精度を得られたことから，意味役割付与に対して提案手法がある程度有効であると考えられる．また，ベースラインの手法は素性を全く考慮しない手法と言い換える事ができるので，この点で提案手法は素性をある程度上手に考慮できていることがうかがえる．

次に表 2 について，「ある状態になる」という概念は単語の「なる」に関連している事が多く，係り受けでのこの概念どうしの組合せは「～になり，～になる」のように用言から用言への係り受けになっていることを意味する．このような場合の深層格の同定には，通常の体言から用言への係り受けに比較して，文脈や常識に関する知識がより重要となる．提案手法はこれらの知識を十分に利用できないため，誤り率が高くなっていると思われる．

表 3 について，表層格「して」は（動詞）＋「して」という形で使用されるので，これに関しても用言どうしの係り受けにおける深層格の同定の困難さが影響していると思われる．

提案手法では係り側・受け側概念と表層格のみを素性としているため，文脈から得られる情報が限定されている．用言どうしの係り受けでの意味役割付与を正確に行うには，文脈情報をより多く投影した素性の選択や，文からは得られない常識に関する知識の利用が必要である．

４．２．出現頻度と精度の関係

データスパースネスと意味役割付与の精度との関係を調べる．係り側・受け側それぞれの概念について学習データでの出現頻度ごとに 10 個の階級に分け，階級の組合せごとにテストデータでの意味役割付与の精度を計算する．階級の分け方はテストデータでの出現回数が約 1,000 件になるように設定している．

結果を表 4 および図 1 に示す．表 4 と図 1 から，係り側・受け側ともに学習データでの出現頻度が高い箇所と低い箇所での精度に大きな差は見られなかった．この結果は非常に興味深い．当初，意味役割付与精度は学習データでの出現頻度が低いほど低く，逆に高くなるほど高くなると考えていたが，この結果はそれを示していない．

この結果が正しいかどうかを調べるために簡単な実験を行った．EDR 電子化辞書には約 40 万概念についての上位下位の階層構造を記述した概念体系辞

表 4. 学習データでの出現頻度別組合せごとの意味役割付与精度（%）

係り側＼受け側	0 ～ 9	10 ～ 29	30 ～ 62	63 ～ 119	120 ～ 197	198 ～ 323	324 ～ 494	495 ～ 828	829 ～ 1520	1521 ～ 28229
0 ～ 6	56.99	63.26	63.18	65.64	67.33	66.92	64.51	64.35	64.63	63.59
7 ～ 20	63.21	64.85	68.62	69.36	68.60	71.28	72.85	64.91	67.39	66.69
21 ～ 42	64.32	67.05	68.47	67.61	67.76	73.77	72.06	68.35	72.38	68.92
43 ～ 78	61.14	65.57	65.16	69.25	70.29	69.01	70.15	68.19	68.96	68.64
79 ～ 127	64.69	70.47	69.19	70.64	67.47	70.70	68.42	71.85	69.05	69.51
128 ～ 202	62.81	66.82	72.00	69.81	67.70	72.01	69.12	67.99	73.11	70.33
203 ～ 309	60.99	66.35	68.90	68.96	67.45	70.20	68.73	70.43	69.14	66.05
310 ～ 480	60.38	68.27	68.48	68.13	70.28	69.52	70.30	69.21	69.05	70.06
481 ～ 943	61.04	70.65	71.62	71.49	70.62	71.71	72.34	70.42	68.36	71.59
944 ～ 14432	56.43	59.97	66.47	64.53	63.67	66.32	62.99	68.57	64.41	61.02

書というものがある．これは概ね木構造となっているので，ある階層値Nを設定し，Nよりも下の階層に属する概念をNに属する上位概念に対応づけることによって概念を抽象化することができる．これにより，データスパースネスを多少回避する事ができる．抽象化した概念を学習データ，調整用データ，テストデータに適用し，再度意味役割付与の実験を行い，抽象化を行っていない場合との精度の比較を行った．その結果，抽象化を行っていない場合と行っている場合の精度に大きな差はなかった．この結果から，学習データでの出現頻度の違いによる精度への影響は多くはないということがわかった．

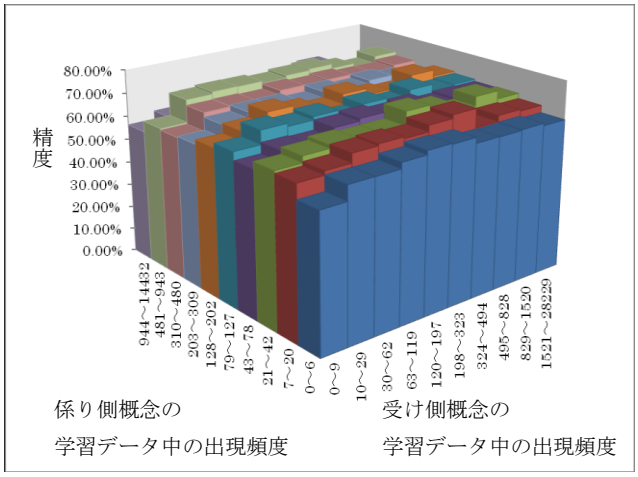

図 1. 学習データでの出現頻度別組合せごとの
意味役割付与精度（%）

５．結論

日本語コーパスを用いたナイーブベイズ法による機械学習により，意味役割付与を行った．評価実験の結果，提案手法にある程度の有用性を確認できた．しかし，用言から用言への係り受けに対しては適切な深層格を付与することができなかった．これは係り側・受け側概念と表層格の情報だけでは文脈に関する情報が得られず，用言どうしの意味的関係を導くに至らなかったからである．これに関しては，文脈の情報をより多く投影した素性の選択や，文からは得られない常識に関する情報の利用を行うことができれば精度を向上させることが可能であると思われる．

また，学習データでの出現頻度と意味役割付与精度の関係を調べた．その結果，学習データでの出現頻度が低い場合と高い場合とで意味役割付与の精度に大きな差は現れないということがわかった．

今後は有用な素性の選択と，統計的手法とルールベースの手法を組み合わせた手法について検討していきたい．

参考文献

[1] 安達昌吾，杉本徹：EDR 電子化辞書を用いた深層格解析手法の改良と評価，第 94 回人工知能学会知識ベースシステム研究会，pp.39-45，2011

[2] 日本電子化辞書研究所：EDR 電子化辞書第二版，2001