機械学習を用いたニ格深層格の自動付与の検討

竹野峻輔† 松田真希子‡ 梶原智之† 山本和英†
長岡技術科学大学電気系† 金沢大学‡
{takeno,kajiwara,yamamoto}@jnlp.org, mts@staff.kanazawa-u.ac.jp

1 はじめに

意味情報を正確に把握する上で，深層格は重要な役割を果たしており，テキスト間の含意関係の判定 [1] など深層格の自動判定技術の応用範囲は広い．特にニ格の深層格推定はデ格に次いで困難であるが，ヲ格に次いでニ格の出現頻度が高い [2] ことから，ニ格の深層格推定の精度向上は重要度が高い．

田辺ら [3] は，ニ格を含む名詞句と係り先の述語の意味属性の付与を行い，ニ格の深層格の推定を行った．その結果，適合率は約 85%，再現率は約 95%をとり，F 値も約 90% に達した．しかし 67 種類もの意味属性を人手で定義したにもかかわらず，オープンテストにおいては F 値が 60%程度しか得られなかった．このことから Rule-based で深層格の付与を行うことには限界がある．

また梅木ら [1] はコーパスに BCCWJ を利用したが，そのコーパスに付与した訓練事例がどの程度の代表性を有し，自動抽出に有効であるかについて検討が必要である．

先行研究において，松田ら [5] は BCCWJ【1】から抽出したニ格を含む文に対し，既存の深層格リストの妥当性について検証を行った．その結果，既存のリストにはない深層格やほとんど出現しない深層格などが見られた．そして深層格の自動付与のため，深層格リストの選定を行う必要があることを示した．

本研究では，表 1 に示されるニ格の深層格リスト [6] について BCCWJ，Web 日本語 N グラム【2】および京都大学テキストコーパス【3】の異なる 3 つのコーパスについて 機械学習によるニ格深層格の自動付与について検討を行う．

2 深層格の自動付与手法

2.1 深層格パターンの機械学習手法

本研究では，深層格の自動付与のため機械学習手法としてナイーブベイズ法を用いた．ナイーブベイズ法は設計が単純でありながら，その分類性能はある程度の高さを持っていることから，汎用性が高く幅広く利用される [7]．ナイーブベイズ法では素性 $f \in F$ の予測深層格 $\hat{c} \in C$ は式 (1) で表される．

各素性の独立性の仮定を行うことで F に対する深層格 c の尤度 $P(c|F)$ は，深層格 c の事前確率 p_c と深層格 c と素性 f の共起確率 $p_{f,c}$ を用いて式 (1) で表される．ただし δ_f は素性 f が入力に含まれるとき 1，含まれないとき 0 をとる関数である．

$$\hat{c} = \arg\max_{c \in C} P(c|f) \tag{1}$$

$$\begin{aligned} P(c|f) &\simeq P(c)P(f|c) \\ &= p_c \prod_{f \in F} \left(p_{f,c}^{\delta_f} (1 - p_{f,c})^{1-\delta_f} \right) \end{aligned} \tag{2}$$

MAP 推定に基づき尤度 $P(c|F)$ を最大化する $p_{f,c}$, p_c は下式で与えられる．$n_{f,c}$ は訓練セット中の素性 f を持つ深層格 c の事例数，n_c は訓練セット中の深層格 c の事例数である．また α は平滑パラメータであり，本研究では $\alpha = 1$ を用いた．

$$p_{f,c} = \frac{n_{f,c} + \alpha}{\sum_{c \in C} n_c + 2\alpha}, \tag{3}$$

$$p_c = \frac{n_c + \alpha}{\sum_{c \in C} n_c + \alpha|C|} \tag{4}$$

2.2 素性の抽出方法

先行研究 [5] では，深層格の分類は前接名詞と後続の動詞・形容詞の組み合わせで意味が決定されることが指摘されている．それに基づき本研究では入力文に対して CaboCha【4】を用いた形態素および係り受け解析を行い，以下に挙げる素性の抽出を行った．ただし係り受け解析で得られる文節リストに対して分類対象とするニ格が含まれる文節を係り元文節，係り元文節の係り先を係り先文節とする．

- 係り元文節に出現する形態素のうち，品詞が記号・フィラー・接頭詞・助詞・助動詞を除いた形態素の原形.
- 係り先文節に出現する形態素のうち，品詞が動詞・形容詞・副詞およびサ変接続可能な名詞である形態素の原形.
- CaboCha の解析結果より得られる，係り元文節および係先文節に付与された固有表現タグ:LOCATION,MONEY,PERSON 等

例として「東京都内にあった家」という文に対して素性抽出を行う．解析結果から，係り元文節と係り先文節はそれぞれ「東京, 都内」と「あった」となる．これより上記の素性抽出を行うと素性は「東京」,「都内」,「ある」に加え，CaboCha の固有表現タグである「LOCATION」が抽出される．

また上記の素性抽出に加え，素性の疎データ問題を解消するため名詞に対しては，日本語語彙大系【5】の意

表 1: ニ格に付与される深層格リスト

深層格	例	定義
1. 時間	8 時に起きる	事象の起こる時間
	同時に起こる	事象・事実の同時関係
2. 場所	近くにある	事象の成立する場所
	東京に行く	事象の主体または対象の最後の位置
3. 結果	二重に折る	修飾関係
	医者になる	変化した後の状態
	息子を医者にする	材料または構成要素
4. 対象	太郎に会う	接触の相手
	父に買ってあげる	利益・不利益の移動先
	対応に怒る	物事が起こる原因，非意志的動作における主体のうち抽象的かつ，発生原因的なもの
	答えに窮する	困難の「行き先」
5. 動作主	太郎に殴られる	有意志動詞を引き起こす主体／受動文の動作主
	太郎に行かせる，行ってもらう	有意志動詞を引き起こす主体／使役表現の動作主
	私には難しい	形容詞文の意味上の主語
6. 目的	映画を見に行く	目的
7. 役割	貿易を外交の手段に用いる	役割・用途 に，として
8. その他	気になる	構成要素の足し算で全体意味が得られないもの等

味カテゴリを用いた汎化を行う．日本語語彙大系中には is-a 関係で結ばれた約 3,000 種類の意味カテゴリが定義されている．本研究では名詞が属するカテゴリをその上位カテゴリに汎化することを名詞の汎化と定義する．

例えば「東京」や「大阪」といった名詞はその上位カテゴリである「都道府県」に汎化できる．

2.3 $tfidf$ を用いた素性選択

訓練データから抽出されたそれぞれの素性についてストップワードの除去を行うため，本研究では素性に対して式 (5) で表される $tfidf$ を計算し素性選択の基準とした．ただし n_c は深層格 c の素性の異なり数，$n_{c,f}$ は訓練データ中における深層格 c の素性 f の出現頻度，df_f は素性 f の出現する深層格の種類数である．

この $tfidf$ は深層格について逆頻度 idf と Harman 正規化された tf との積で表され，限られた深層格で高頻度で出現する素性であるほど高い値をとる．

$$tf_{c,f} = \frac{\log_2 n_{c,f}}{\log_2 n_c},$$
$$idf_f = \log_2 \left(\frac{|C|}{df_f} \right) + 1,$$

$$tfidf_{c,f} = tf_{c,f}\, idf_f \tag{5}$$

3 評価実験および考察

3.1 訓練および評価セットとその評価方法

本研究では深層格が付与された訓練セットとして BCCWJ より 25,936 件，京都大学テキストコーパスより 10,711 件，Web 日本語 N グラムより 7-gram データを高頻度順に 9,828 件用意した．また訓練セットとは別に様々な文書[1] から文の抽出を行い，評価用コーパスとして 1,037 件を用意した．

本研究では深層格の全体の分類精度の評価指標として平均正答率を式 (7) に，深層格別の評価指標として適合率 P_c, 再現率 R_c，F_c を式 (6) に定義する．ただし n_{right} は深層格 c について分類結果のうち正解した事例数, n_c は分類結果が深層格 c となる事例数，n_t は評価セット中の深層格 c の事例数である．

$$P_c = \frac{n_{right}}{n_c}, \quad R_c = \frac{n_{right}}{n_t},$$
$$F_c = \frac{2\, P_c\, R_c}{P_c + R_c} \tag{6}$$

$$S_{ave} = \frac{\sum_{c \in C} n_{right}}{\sum_{c \in C} n_c} \tag{7}$$

3.2 汎化の深さの変化に対する自動付与結果の評価

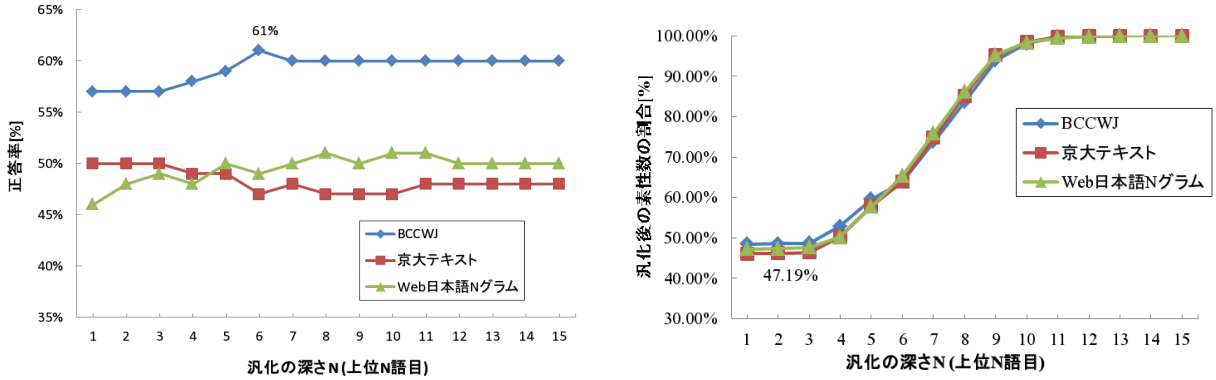
本研究では，意味カテゴリの抽象度が正答率にどの程度影響するかを調査するため，根から汎化先となる意味カテゴリまでの距離を汎化の深さ d_n として定義する．例として「東京」「大阪」の上位の意味カテゴリである「都道府県」は「固有名詞⇒地名⇒地域名⇒行政区画名⇒日本⇒都道府県」というように根から 6 つ目の意味カテゴリであり，汎化の深さ d_n は 6 となる．

汎化の深さを 1〜15 まで変化させたときの正答率および訓練セット中の素性数の変化を測定した．図 1 にその結果を示す．

図 1a より，最大の正答率は汎化の深さ 6 に対して BCCWJ の 61%であるが，全体的に汎化による正答率への影響は小さい．

一方で図 1b に着目すると，汎化深さに応じて素性数には大きな変化が見られ，コーパスに依らず同様の減少傾向を示す．汎化前に比べ，最大で 47.19%まで素性数が減少している．

[1]出典:青空文庫, 読売新聞, アメーバブログ, Wikipedia, Twitter, 論文, 論文抄録, サイゾーウーマン, サイゾー



(a) コーパスごとの正答率の変化

(b) コーパスごとの素性の種類数の変化

図 1: 汎化深さに対する各コーパスの自動付与結果の比較

これより汎化による素性数削減の有効性が確認できた．

3.3 $tfidf$ の素性選択に対する深層格自動付与の評価

$tfidf$ についてある閾値を設け，閾値以下の素性を取り除くことで，正答率の改善が見られるか確認するため測定を行った．

BCCWJ に対し汎化の深さを 6 とし，$tfidf$ の閾値を 0.0 から 1.5 まで変化させた時の測定結果を図 2 に示す．また各深層格について $tfidf$ が高かった上位 5 つの素性を表 2 に示す．

図 2a より閾値を高くすることでの大きな正答率の改善は見られない．また素性数については図 2b より深層格ごとの差異も小さい．なお閾値に対して指数的な減少傾向にある．正答率が最も高かった閾値 0.4 において素性数は 37%まで減少している．

表 2 の素性を参照すれば深層格「1. 時間」の「同時」や「DATE」，「6. 目的」の「ため」など人間の感覚的にも近い素性が上位に表れていることがわかる．しかしながら上位に出現する素性の多くは動詞，形容詞または副詞であり，汎化された名詞は「気」を除き上位に出現していない．図 1 において汎化の深さに対し正答率が変化していないことからも，汎化された名詞は，深層格の推定結果にさほど反映されていないことがわかる．

これより $tfidf$ により深層格に関連性の高い素性が選択可能であり，閾値を設けることで素性数の削減を行うことができた．しかし $tfidf$ の閾値による素性選択では分類器の正答率の向上には至らなかった．

図 2a から正答率に対して大きな重みを占める深層格は「8. その他」と「4. 対象」である．今回用いた 3 種類のコーパスに関していずれも，この 2 つの深層格の重みが高く，事例数はこの 2 つの深層格をあわせると総事例数の 70%以上を占めていることがわかった．従ってこれらの深層格の推定に対して，改善を施すことが深層格の付与性能向上のため肝要である．

さらに 2 つの深層格に対して，結果を分析したところ「8. その他」に関しては分類器による再現率は平均 37%，精度が平均 87%と再現率が低い．一方で「4. 対象」に関しては再現率は平均 90%，精度は平均 46%と精度が低くなっていることが分かった．

「8. その他」では「気になる」や「耳にする」といった固有表現が多いことから，素性に対して一貫性が少ない．また高頻度で表れる素性も「なる」,「する」など他の深層格でも高頻度に出現するものが多い.「8. その他」の事例数は多いことから閾値による素性選択でも除去される素性も比較的多くなる．

これらの素性は $tfidf$ の閾値の素性選択により取り除かれやすいため，結果再現率の低下に繋がっている．従ってこれを解消するためには訓練セットの事例数を多く学習させること，つまり「8. その他」について $tfidf$ の閾値を低くすることで素性数が保たれ再現率の向上が期待できる．

また「4. 対象」は多くの深層格の推定の不正解結果として出力されていることから精度が大きく低下していた．これは本研究では，素性に関して独立性を仮定したことから，動詞と名詞の組み合わせによる深層格の判別，例えば同一の動詞で深層格が異なる場合などを分類器に反映させることができておらず，精度低下を招いている．従ってこれらの組み合わせを考慮した素性抽出を行うことで「4. 対象」の精度向上が期待できる．

3.4 各コーパスの比較

本研究では 3 種類のコーパスについて，それぞれを訓練または評価セットとして測定を行い，それぞれのコーパスが訓練セットとして妥当であるか評価を行った．各コーパスの事例数による差異を無くすため，訓練セットのコーパスから 8,500 件，評価セットのコーパスから 1,037 件を無作為に取り出し測定を行った．

各コーパスの組み合わせについて $tfidf$ の閾値 0.0〜1.5，汎化の深さをは 0 〜 15 まででグリッドサーチを行い最大正答率を求めた．各コーパスの最大正答率の比較結果を表 3 に示す．

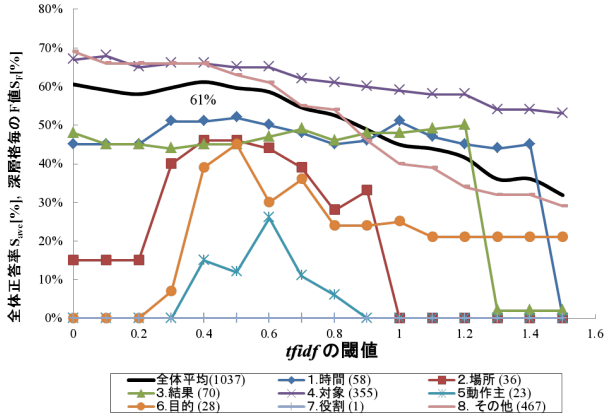
Web 日本語 N グラムは $tfidf$ の閾値 0.4，汎化の深さ 12 であるとき，自身を評価セットに用いた時に最大正答率 84%と高い正答率を記録している．しかしながらその他のコーパスを評価セットとした場合の正答率は低い．よって未知のデータに対し 3 つのコーパスの中で Web 日本語 N グラムを用いるのは適切でない．

一方 BCCWJ は Web 日本語 N グラムと同様，自身を評価セットに用いた場合で正答率が 70%と最も良い．また未知のデータである評価用コーパスに対しても 3 つのコーパス中では BCCWJ が $tfidf$ の閾値 0.5, 汎化の深さ 5 において最大正答率 62%を得た．

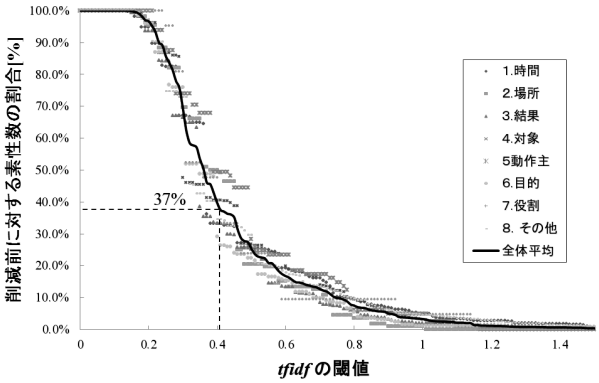
これより BCCWJ，京都大学テキストコーパス，Web 日本語 N グラムの 3 つの内では BCCWJ が訓練セットとして優れていることがわかった．

4 おわりに

本研究では BCCWJ，Web 日本語 N グラム，京都大学テキストコーパスの 3 つのコーパスを利用し，ナイー



(a) 各深層格の F 値および正答率の変化



(b) 各深層格および全体の素性数の変化

図 2: $tfidf$ に対する深層格の自動付与結果 (BCCWJ)

表 2: $tfidf$ の高い素性 上位 5 項目 ※ () 内は素性の $tfidf$ を表す

1. 時間	2. 場所	3. 結果	4. 対象	5. 動作主	6. 目的	7. 役割	8. その他
同時 (1.66)	部屋 (1.33)	こと (1.51)	基づく (1.32)	私 (1.07)	ため (1.52)	間食 (1.41)	対す (2.08)
DATE(1.40)	どこ (1.15)	至る (1.33)	含む (1.24)	れる (0.93)	散歩 (1.03)	みやげ (0.81)	対する (1.98)
すぐ (1.35)	そこ (1.12)	なる (1.28)	船 (1.21)	誰 (0.91)	の (0.96)	温める (0.71)	関する (1.62)
年 (1.34)	LOCATION(1.00)	倍 (1.17)	そこ (1.17)	れ (0.88)	出る (0.94)	記念 (0.68)	最後 (1.59)
即座 (1.21)	住む (0.99)	**気** (1.17)	こと (1.16)	千賀 (0.82)	助け (0.89)	する (0.67)	次 (1.56)

表 3: コーパス同士の比較

評価＼訓練	C1	C2	C3
C1	70%	57%	58%
C2	45%	49%	44%
C3	63%	60%	84%
評価用コーパス	62%	53%	54%

C1：BCCWJ
C2：京都大学テキストコーパス
C3：Web 日本語 N グラム

ブベイズ法によるニ格深層格の自動付与の検討を行った．その際に日本語語彙大系を利用した汎化および $tfidf$ による素性選択を試みたが，その結果として分類性能の大きな向上は見られなかった．しかしながら表 1 のニ格の深層格推定においては「8. その他」,「4. 対象」が大きな重みを占めていることが判明した．

またコーパスの訓練セットとしての妥当性を検討したところ，3 つのコーパスのうち未知のデータである評価用コーパスに対して BCCWJ を用いた場合が最も良く，自動付与の正答率は最大 62%が得られた．

今後の課題としては深層格「8. その他」と「4. 対象」に合わせた改良を行い，自動付与性能の向上を目指すこととする．

謝辞

本研究は科学研究費補助金基盤研究（B）[課題番号 23320105] の助成を受けて行われた．

利用した言語資源およびツール

【1】国立国語研究所. 現代日本語書き言葉均衡コーパス (BCCWJ). 国立国語研究所, 2011. http://www.ninjal.ac.jp/corpus_center/bccwj.

【2】工藤拓, 賀沢秀人. Web 日本語 N グラム第 1 版, 言語資源協会, 2007. http://www.gsk.or.jp/catalog/gsk2007-c/.

【3】黒橋禎夫, 河原大輔. 京都大学テキストコーパス・プロジェクト. 言語処理学会 第 3 回年次大会, pp.115-118, 1997.

【4】工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp.1834-1842, 2002.

【5】白井諭, 大山芳史, 池原悟, 宮崎正弘, 横尾昭男. 日本語語彙大系について. 情報処理研究報告.IM, Vol.98, No.106, pp.47-52, 1998.

参考文献

[1] 梅基宏, 杉原大悟, 大熊智子, 増市博. LFG 解析と語彙資源を利用した日本語含意関係判定. 自然言語処理研究会報告, No.113, pp. 57–64, 2008.

[2] 渋木英潔, 荒木健治, 桃内佳雄, 栃内香次. 単語概念の深層格選好に基づく深層格推測手法. 電子情報通信学会論文誌, J89-D, No.6, pp. 1413–1428, 2006.

[3] 田辺利文, 吉村賢治, 首藤公昭. 格助詞「に」の深層格推定 - 格助詞の意味再考 -. 情報処理学会研究報告, No.113, pp. 65–72, 2009.

[4] 奥田靖雄. に格の名詞と動詞とのくみあわせ. 日本語文法・連語論 (資料編), pp. 281–323, 1983.

[5] 松田真希子, 森篤嗣, 川村よし子, 庵功雄, 山口昌也, 山本和英. 日本語深層格の自動抽出のためのコーパス開発. 言語処理学会第 18 回年次大会発表論文集, pp. 205–208, 2012.

[6] 松田真希子, 森篤嗣, 川村よし子, 庵功雄, 山本和英, 山口昌也. ニ格深層格の定量的分析. 言語処理学会第 20 回年次大会発表論文集, 2014.

[7] 岩沢拓未, 杉本徹. 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. 言語処理学会第 18 回年次大会発表論文集, pp. 386–389, 2013.