深層格選好に基づく深層格推測手法の英文への適用

渋木 英潔 †◇　荒木 健治 ‡　桃内 佳雄 *◇　栃内 香次 †

† 北海学園大学大学院経営学研究科　‡ 北海道大学大学院情報科学研究科
* 北海学園大学工学部　◇ 北海学園大学ハイテク・リサーチ・センター

1　まえがき

我々は，人手による労力の軽減を目的として，深層格を自動的に推測する手法の研究を行っている [1, 2]. 深層格を推測するための知識は，多様な言語表現に対処するため，深層格のタグが付与されたデータを用いて学習されることが多い [3, 4]. しかしながら，タグの付与には多大な労力が必要であり，可能な限りタグを付与せずに学習できることが望ましい．このような背景から，我々の手法では，一定量のタグ付きデータと大量のタグなしデータから学習することを試みている．本手法は，単語の概念ごとに，それぞれ特定の深層格に解釈される傾向をもつという仮定に基づき，タグ付きデータから深層格の傾向を計算した後，その傾向を用いてタグなしデータから多様な言語表現に対処するための知識を学習する．これにより，人手による労力の軽減を行う．この深層格の傾向を，本稿では深層格選好と呼び，深層格選好に基づいて深層格を推測する本手法を DCAPR(**D**eep **C**ase **A**nalysis based on **P**reference of deep case and **R**egularization) と呼ぶ. 先行研究 [1, 2] では日本語を対象として実験を行っていたが，深層格選好はどの言語にも存在すると考えられるため，本手法は日本語以外の言語にも適用可能であるといえる．

本稿では本手法を英文へ適用することを試み，適用するための変更箇所を説明した後，EDR 英語コーパス [5] を利用した実験の結果について報告する．

2　全体の流れ

本手法の全体の流れを図 1 に示す．学習データとテストデータは共に構文解析済みの単文である．最初に，

学習データとテストデータで使用される名詞と動詞のクラスタリングを，字面や語順などの表層的な情報に基づいて行う．表層的な情報に基づくのは，人手による労力を軽減するため，機械的に処理できる情報に限定したためである．次に，　作成した単語クラスタ

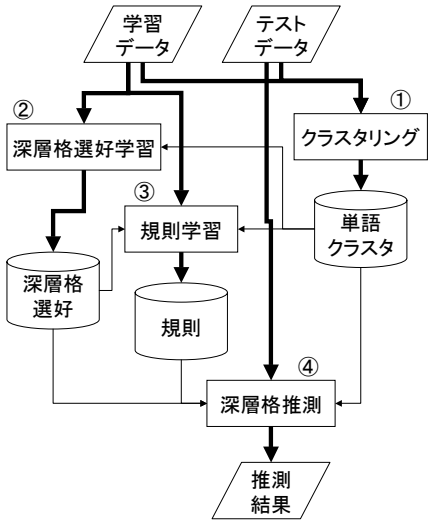

図 1: 全体の流れ

を参照し，単語及び単語クラスタの深層格選好を学習データに付与された深層格タグに基づいて学習する．深層格選好は，ある深層格と解釈される傾向を 0 から 1 の範囲で表現したものであり，深層格選好が大きな値であるほど，その深層格と解釈される傾向にあることを意味する．その後，　学習された深層格選好を基に，学習データの深層格を推測し，推測された深層格と学習データ中の表層表現とを対応付ける規則を学習する．本来であれば，深層格選好の学習に用いられるタグ付きデータとは別に，規則を学習するためのタグなしデータを用意することが望ましいが，本稿では用意可能だったデータ量の関係から，深層格選好学習に用いたデータからタグを除いたものを規則学習のデータとした．最後に，　作成されたクラスタ及び学習された深層格選好と規則に基づいて，テストデータの深層格を推測する．

3　知識表現

深層格選好の例を表 1 に示す．名詞の深層格選好は，ある動詞に下位範疇化される際に，その名詞が担う深

表 1: 深層格選好の例

NOUN	agent	object	goal	condition	···	quantity
I	1.0	0.0	0.0	0.0	···	0.0
Tokyo	0.0	0.0	0.4	0.0	···	0.0
VERB	agent,agent	agent,object	agent,goal	agent,condition	···	quantity,quantity
live	0.0	0.0	0.0	0.0	···	0.0
give	0.0	0.7	0.4	0.0	···	0.0

表 3: 規則の例

RULE	agent	object	goal	condition	···	quantity
[ϕ:F1][in:B1][1]	1.0	0.0	0.0	0.0	···	0.0
[ϕ:B1][to:B2][2]	0.0	0.0	0.8	0.0	···	0.0

表 2: 深層格の一覧

agent	object	goal
condition	implement	material
place	scene	source
cause	purpose	basis
beneficiary	quantity	

層格の傾向を表し，動詞の深層格選好は，ある名詞を下位範疇化する際に，その名詞が担う深層格の傾向を表す．また，動詞は複数の名詞を下位範疇化するため，動詞の深層格選好は一つの深層格ではなく，下位範疇化する名詞群が担う深層格の組として表現する．本稿では，二つの名詞と一つの動詞からなる文を対象としているため，動詞の深層格は表 1 に示すように二つの深層格の組に対して表現される．対象となる深層格は，大石ら [3] と小山ら [4] でそれぞれ対象とした深層格を併せ，表 2 に示す 14 種類とした[1]．

規則は，文法語や語順などに代表される，ある言語ごとに存在する深層格を推測するための情報を利用するための知識である．名詞の深層格選好と同様に，規則を適用した場合に解釈される深層格の傾向を 0 から 1 の範囲で表現する．日本語を対象とした先行研究 [2] では，名詞に後続する助詞，名詞の文中の位置，文に含まれる全ての助詞を一つのパターンとして表現したもの[2]の三つの情報を規則を適用する際の条件としていた．英語では前置詞が日本語の助詞に相当すると考えられるため，助詞の代わりに前置詞を用いることとし，文法語パターンも前置詞を基に作成することとした．ただし，日本語では助詞が省略可能であっても基本的に存在するのに対し，英語では前置詞が存在しないことがあるため，空の字面 ϕ も他の前置詞と同様に扱うこととした．語順に関して，日本語では基本的に動詞の前方に位置することから文頭からの順序で表現したのに対し，英語では動詞の前後に位置するため，動詞からの相対的な方向と距離で表現することとした．動詞の前方に位置する場合には F，後方に位置する場合には B の記号を用い，記号の後ろに動詞からの距離を併記して語順を表す．F1 は動詞の直前の位置を意味し，B2 は動詞の二つ後ろの位置を意味する．規則の例を表 3 に示す．三つの [ ] で括られた規則の適用条件のうち，最初の二つは，下位範疇化される二つの名詞の文法語と語順の情報である．また，二つ一組で文法語パターンも表している．最後の [ ] は，規則を適用する名詞が文法語パターンで下位範疇化されている名詞のどちらであるかを示す情報である．表 3 の最初の規則は，例えば，“I live in Tokyo.” において “I” の深層格を推測するために利用され，次の規則は “Bring it to me.” において “me” の深層格を推測するために利用される．

4 学習

クラスタリングの方法は，先行研究 [2] に記述しており，英文に適用する際の変更箇所は，前章で述べた文法語と語順のみであるため，紙面の都合上，本稿では省略する．

[1]深層格の種類は，目的とするタスクの種類に応じて定義されるものと考えており，表 2 の深層格が最善であるとは考えていない．本稿では，日英における本手法の比較を行うために，EDR の概念関係子に基づくこととした．

[2]本稿では，文法語パターンと呼ぶ．

4.1 深層格選好学習

以下の手順で深層格選好を学習する．ある名詞 n が学習データ中で付与されている深層格 d の頻度を $fq_N(n,d)$ とする．以下の式 (1) に従って 0 から 1 の範囲となるよう深層格選好 $dp_N(n,d)$ を計算する．

$$dp_N(n,d) = \frac{fq_N(n,d)}{\sum_{i \in D} fq_N(n,i)^2} \quad (1)$$

D は深層格の全集合である．これは各要素の頻度をベクトル要素とした場合の単位ベクトルに相当する．$0 \leqq dp_N \leqq 1$ となるよう深層格選好を計算するためには，他にも確率として表現する方法が考えられるが，予備実験においてベクトル表現とした方が良い結果であったためベクトル表現とした．

動詞 v の場合も同様に，二つの深層格 d_1，d_2 の頻度を $fq_V(v,d_1,d_2)$ とし，以下の式 (2) に従って深層格選好 $dp_V(v,d_1,d_2)$ を計算する．

$$dp_V(v,d_1,d_2) = \frac{fq_V(v,d_1,d_2)}{\sum_{i,j \in D} fq_V(v,i,j)^2} \quad (2)$$

また，作成した名詞クラスタ nc 及び動詞クラスタ vc の深層格選好 $dp_{NC}(nc,d)$ 及び $dp_{VC}(vc,d_1,d_2)$ を，以下の式 (3) 及び式 (4) に基づいて $fq_{NC}(nc,d)$ 及び $fq_{VC}(vc,d_1,d_2)$ を求めた後，同様にして計算する．

$$fq_{NC}(nc,d) = \sum_{i \in nc} fq_N(i,d) \quad (3)$$

$$fq_{VC}(vc,d_1,d_2) = \sum_{i \in vc} fq_V(i,d_1,d_2) \quad (4)$$

4.2 規則学習

最初に，学習した深層格選好を用いて，学習データの深層格を推測する．動詞 v に下位範疇化されている名詞 n が深層格 d と推測される尤度 $pl_L(n,d)$ は以下の式 (5) により計算される．

$$val_L(n,v,d) = dp_N(n,d) \times \sum_{i \in D} dp_V(v,d,i)$$

$$pl_L(n,v,d) = \frac{val_L(n,d)}{\sum_{i,\in D} val_L(n,i)^2} \quad (5)$$

名詞と動詞の深層格選好の積 $val_L(n,v,d)$ を求め，その値を深層格選好と同様に 0 から 1 の範囲に変換した値が尤度となる．尤度が閾値 T 以上となった深層格を学習データにおける推測結果とする．このとき，複数の深層格の尤度が閾値以上となる場合や，どの深層格の尤度も閾値未満となる可能性があるため，推測結果が一意に定まるとは限らない．

次に，ある動詞に下位範疇化されている全ての名詞の深層格が一意に推測され，かつ，推測された深層格が一文一格の原理を満たしている文を抽出する．この条件を満たす文の推測結果は，正解である可能性が高いため，これらの文を用いて規則の学習を行う．名詞 n に付属する前置詞や語順や文法語パターンの情報を g とし，深層格 d が推測された頻度 $fq_R(g,d)$ を求める．深層格選好と同様に 0 から 1 の範囲となるよう以下の式 (6) に基づいて計算し，前置詞などの情報 g を満たす名詞が d となる傾向を示す規則 $gr(g,d)$ を学習する．

$$gr(g,d) = \frac{fq_R(g,d)}{\sum_{i \in D} fq_R(g,i)^2} \quad (6)$$

5 深層格推測

テストデータの深層格は，深層格選好と規則に基づいて推測される．名詞 n が前置詞などの情報 g を伴って動詞 v に下位範疇化されている場合，n が深層格 d と推測される尤度 $pl_A(n,v,g,d)$ は以下の式 (7) により計算される．

$$val_A(n,v,g,d) = val_L(n,v,d) \times gr(g,d)$$

$$pl_A(n,v,g,d) = \frac{val_A(n,v,g,d)}{\sum_{i,\in D} val_A(n,v,g,i)^2} \quad (7)$$

ただし，n や v が学習データに存在しない場合には，$dp_N(n,d)$ や $dp_V(v,d_1,d_2)$ の代わりに $dp_{NC}(nc,d)$ や $dp_{VC}(vc,d_1,d_2)$ を用いて $val_L(n,v,d)$ を求める．閾値 T 以上の尤度となった深層格が推測結果となる．

6 実験

EDR 英語コーパスから，二つの名詞を下位範疇化する能動態の単文であり，文を構成する名詞，動詞，前置詞の出現頻度が三回以上である 7,485 文を抽出した．テストデータとしてランダムに 200 文を選択し，残りの 7,285 文を学習データとした．テストデータの推測結果を，以下の式に従って再現率と精度により評価した．

$$\text{再現率} = \frac{\text{正解出力数}}{\text{正解数}}. \quad \text{精度} = \frac{\text{正解出力数}}{\text{出力数}}.$$

$$F \text{ 値} = \frac{2 \times \text{再現率} \times \text{精度}}{\text{再現率} + \text{精度}}.$$

閾値 T は，これまで和文の深層格を推測する際に用いた閾値と同じ 0.7 とした．

結果を表 4 に示す．一行目が深層格選好と規則を用いた本手法の結果である．二行目と三行目は，深層格

表 4: 深層格の一覧

		評価名詞数	出力数	正解数	精度	再現率	F 値
英文	提案手法	400	394	297	75.4%	74.3%	74.8%
	深層格選好のみ	400	393	286	72.8%	71.5%	72.1%
	規則のみ	400	400	146	36.5%	36.5%	36.5%
和文	（文献 [2] からの引用）	380	450	337	74.9%	88.7%	81.2%

選好または規則のみを用いてテストデータを解析した結果であり，本手法の学習が有効であることを比較調査するために行ったものである．四行目は EDR 日本語コーパスを用いた先行研究 [2] の結果であり，和文との比較を行うために引用した．

深層格選好は学習データ中の統計情報に基づいており，深層格選好のみを用いた推測は統計的手法の一種とみなすことができる．一行目と二行目を比較すると，学習された規則が，深層格選好のみを用いた推測結果を，精度，再現率共に向上させていることが確認できる．しかしながら，三行目の結果は，規則の推測精度自体は高いものではないことを示している．一文一格の原理に基づいて学習された規則は，深層格選好のみでは正しく推測できない文を，正しく推測できるように補助していると考えられる．

和文を用いた場合と比較を行うと，英文の方が出力数が抑えられており，結果として再現率が低下し精度が向上している．F 値を比較すると英文の方が低い値であるが，この差は再現率低下の影響によるところが大きい．また，字面に基づいて受動態を識別したために，不規則変化動詞による受動態を完全に除外することができなかったことによる学習及び推測への影響も挙げられる．したがって，データの見直しを行った後，閾値を下げ出力数を増加させるなど，最適な閾値の調査を今後行っていきたい．

精度のみを比較した場合，和文と英文との間に大きな差は存在せず，本手法が英文においてもある程度有効であると考えられる．ただし，精度の値を絶対的に評価した場合には十分な値とはいえず，今後改善する必要がある．

7　まとめ

我々は，これまで一定量のタグ付きコーパスから深層格選好を単語の概念ごとに一旦計算し，それに基づいて深層格を自動的に推測する手法を提案してきた．本稿では，提案手法を英文へ適用することを行い，本手法が日本語以外の言語でも有効であるかどうかを判断した．英文へ適用するにあたり，変更箇所は，文法語や語順といった表層情報の再定義に留め，基本的な枠組みは日本語の場合と同じ枠組みとした．EDR 英語コーパスを用いて実験した結果，再現率が 74.3%，精度が 75.4%となり，日本語コーパスを用いた場合の精度 74.9%（再現率 88.7%）と比較して，顕著な差が表れなかった．このことから，本手法が英語においても有効であることが確認でき，特定の言語に依存しない可能性が示唆された．今後は，精度の向上を図ると共に，機械翻訳などの多言語を扱う手法の補助として本手法を応用していきたいと考えている．

謝辞

本研究の一部は，北海学園大学ハイテク・リサーチ・センター研究費による補助のもとに行なわれた．

参考文献

[1] 渋木英潔, 荒木健治, 栃内香次: 一文一格の原理と規則化に基づいた深層格の自動推測手法, FIT2003 情報科学技術フォーラム情報技術レターズ, pp.91–92, (2003).

[2] 渋木英潔, 荒木健治, 桃内佳雄, 栃内香次: 深層格の推測手法における自動クラスタリングの利用, FIT2004 情報科学技術フォーラム情報技術レターズ, pp.79–80, (2004).

[3] 大石亨, 松本裕治: 格パターン分析に基づく動詞の語彙知識獲得, 情報処理学会論文誌, Vol.36, No.11, pp.2597–2610 (1995).

[4] 小山正太, 乾伸雄, 小谷善行: 「名詞と表層格」パターンに対する深層格対応の推測, 情報処理学会研究報告, NL-154-22, (2003).

[5] （株）日本電子化辞書研究所: EDR 電子化辞書使用説明書 (1995).