LF-002

深層格の推測手法における自動クラスタリングの利用
Use of Automatic Clustering in a Method for Inference of Deep Cases

渋木 英潔 †* 荒木 健治 ‡ 桃内 佳雄 ††* 栃内 香次 ‡‡
Hideyuki Shibuki Kenji Araki Yoshio Momouchi Koji Tochinai

1. まえがき

我々は，人手による労力を軽減することを目的とした深層格の自動推測手法の研究を進めてきた [1, 2]．我々の手法は，以下の考え方に基づいて行われる．単語の概念ごとに深層格の選好[1]を数値化して表現し，選好値を手がかりとして，助詞などの表層表現と深層格との対応関係を発見する．その後，発見された対応関係を規則として学習し，規則に基づいて再解析することで推測精度の向上を行う．従って，深層格選好が類似した傾向にある単語をまとめ，選好値を計算しておくことが必要となる．これまでの我々の手法では，類似した単語を調べるために分類語彙表を利用しており，未登録語に対する課題が残されていた．また，従来手法である大石ら [3] においても分類語彙表を利用しており，同様の問題が考えられる．このような背景から，本稿では，人手で作成されたシソーラスに依存しない手法を提案する．提案手法では，助詞や語順という表層的な情報に基づいてクラスタリングを行い，その結果を分類語彙表の代わりに使用する．これにより，多大な労力をかけることなく，未登録語の問題を改善し，深層格推測の精度を向上させる．

次節では表層情報について述べた後，3. 節でクラスタリングの方法と結果について述べる．4. 節では，作成されたクラスタを用いて深層格を推測した結果について述べる．5. 節は結論である．

2. 表層情報

これまで，Caraballo[4] など，表層的な情報を手がかりに，意味的に類似した単語のクラスタリングを行う手法が提案されている．本稿では，深層格選好が助詞の出現頻度及び語順に反映されていると仮定し，これらの情報に基づいて，深層格選好の傾向が類似した単語のクラスタリングを試みる．また，人手による労力を軽減しつつ，単語の被覆率を向上させることが目的であるため，可能な限り字面や語順などの自動抽出可能な情報を手がかりとすることとした．

名詞の深層格選好を反映する表層情報として，文全体の助詞パターンと助詞の位置を用いる．本稿での助詞パターンは，助詞の順序を考慮しており，「太郎が本を読む」と「本を太郎が読む」では別の助詞パターン (が, を) 及び (を, が) とする．また，「8 時に札幌に行く」のように，同じ助詞が連続する場合が考えられるため，どの位置の助詞であるかの情報も必要である．従って，「太郎が本を読む」における「太郎」の手がかりとなる表層情報は，(が, を:1) のように表現される．

表 1: 表層情報と深層格情報の例

動詞

	(が, を)	(が, に)	(を, が)
	[agent:object]	[agent:goal]	[object:goal]
見る	6	0	1
	9	0	0
行く	0	8	0
	0	8	2

名詞

	(が, を:1)	(が, を:2)	(が, に:1)
	[agent]	[object]	[goal]
太郎	8	4	7
	10	6	4
東京	0	2	2
	0	3	6

動詞の選好値は，個々の深層格ではなく，深層格の組を単位として表現する．例えば，「太郎が [agent] 花子を [object] 見た」と「東京で [place] パンダを [object] 見た」において，「見る」の選好値を，[agent] が 1 回，[object] が 2 回，[place] が 1 回とするのではなく，[agent:object] が 1 回，[object:place] が 1 回というように表現する[2]．この深層格の組を深層格パターンと定義する．動詞は深層格パターンの類似性に基づくクラスタリングを行うため，深層格パターンを反映する表層情報として助詞パターンのみを用いることとした．

3. クラスタリング

本稿では，閾値以上の類似度をもつ単語群を同一のクラスタにまとめる．単語間の類似度は，Caraballo[4] と同様に，2 つの単語ベクトルの余弦で定義する:

$$\text{類似度}\ (\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{w}}{|\vec{v}|\ |\vec{w}|}.$$

単語ベクトルの要素は，その単語と共に出現した表層情報の頻度である．動詞と名詞の表層情報及び深層格情報に関する例を表 1 に示す．各単語の上段が表層情報，下段が深層格情報である．

†北海学園大学大学院経営学研究科, Graduate School of Business Administration, Hokkai-Gakuen University
‡北海道大学大学院情報科学研究科, Graduate School of Information Science and Technology, Hokkaido University
††北海学園大学工学部, Faculty of Engineering, Hokkai-Gakuen University
‡‡北海学園大学経営学部, Faculty of Business Administration, Hokkai-Gakuen University
*北海学園大学ハイテク・リサーチ・センター, High-Tech Research Center, Hokkai-Gakuen University

[1]本稿での深層格の選好とは，例えば，「太郎」，「花子」，「東京」，「札幌」の 4 つの名詞を比較した場合，「太郎」や「花子」は「東京」や「札幌」よりも agent と解釈される傾向が高く，「東京」や「札幌」は place と解釈される傾向が高い，ということである．また，動詞との関係においても，例えば，「車」と「見る」の間の関係として，「車を見る」という object としての解釈の方が，「車の中で何かを見る」という place としての解釈よりも好まれやすい，ということが挙げられる．

[2]深層格の組で表現する理由は，後の深層格推測において，以下のような効果を期待するからである．上の例を基に選好値を求め，「札幌で雪を見た」の推測を行うとする．「札幌」の深層格が place と推測できた場合，深層格単位で表現されていたとするならば，「雪」の深層格として agent と object が考えられるのに対して，深層格の組による表現ならば一意に object と決定することができる．

表 2: クラスタリング結果

	動詞				名詞			
閾値	クラスタ数	平均 DS	分散	調査単語数	クラスタ数	平均 DS	分散	調査単語数
0.5	12	0.83	0.13	229	15	0.83	0.17	536
0.6	19	0.83	0.11	228	28	0.85	0.14	535
0.7	38	0.86	0.10	217	61	0.88	0.12	529
0.8	86	0.90	0.08	191	141	0.91	0.09	485
0.9	157	0.94	0.07	104	355	0.95	0.06	261
分類語彙表	37	0.80	0.12	16	182	0.88	0.10	187

表層情報に基づいて作成されたクラスタが，分類語彙表と比較して，どの程度深層格選好の類似性を表現できるかを検討するために EDR コーパスを用いた調査を行った．調査に用いた文は以下の手順により求めた．EDR コーパスから推測対象となる深層格[3]をもつ意味フレームに対応する動詞と名詞の係り受け関係を抜き出した．次に，格に変化を及ぼす受身と使役の助動詞を含む事例を除外した．その後，2 つの名詞を下位範疇化している動詞を事例として抽出し[4]，その中から，事例を構成する，全ての名詞，動詞，助詞の使用頻度が 3 回以上である 1,900 事例を対象に調査を行った．対象データに含まれる動詞数は 229，名詞数は 536 であった．また，単語間の類似性を調査するという目的から，複数の単語がクラスタリングされているものを調査対象とした．

閾値を 0.5 から 0.9 まで変化させた場合，ならびに，分類語彙表の分類に従った場合の結果を表 2 に示す．深層類似度 (DS) とは深層格情報をベクトル要素とした場合の類似度であり，表中の平均 DS は，所属するクラスタの重心との深層類似度を平均した値である．分散は深層類似度における重心との二乗誤差平均である．深層格選好が表層情報に反映されているという仮定が正しいならば，閾値を高くするに従い，平均 DS は高く，分散は小さくなると考えられる．表 2 から，この仮定が正しかったと考えられ，閾値を 0.8 以上とした場合，分類語彙表を用いる場合よりも，名詞，動詞共に深層格選好の傾向が類似した単語をクラスタリングすることができた．また，調査単語数の比較から，いずれの閾値においても単語の被覆率が向上しており，未登録語の問題を改善することができた．

4. 深層格推測実験

分類語彙表よりも高い平均 DS を示すクラスタを用いることで，先行研究 [2] よりも深層格推測の精度を向上させることができると考えられる．深層格の推測手法は，先行研究とほぼ同一であり，主な変更箇所として，1) 動詞の選好値を深層格パターン単位とした，2) 助詞パターンにおける助詞の順序を考慮した，3) 選好値を確率表現からベクトル表現とした，4) 正解の可能性のある深層格を全て出力することとした，が挙げられる．これに伴い，評価基準を再現率と精度とし，それぞれ以下の式により計算した．

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

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

正誤の判断は EDR コーパスの概念関係子と一致したかどうかによった．3. 節で作成した 1,900 事例の内，ランダムに選択した 190 事例を評価データとし，残りの 1,710 事例を学習データとした．分類語彙表に代えて用いるクラスタは，平均 DS と評価単語数の被覆率から閾値を 0.8 とした場合の結果とした．

表 3: 深層格推測実験結果

正解数	380	再現率	88.7%
出力数	450	精度	74.9%
正解出力数	337	F 値	81.2%

結果を表 3 に示す．分類語彙表を用いた先行研究 [2] では精度が 61.6%であり，今回の実験では 13.3 ポイント向上した 74.9%であった．このことから，深層格推測手法において，表層表現に基づくクラスタを分類語彙表に代えて用いることの有効性を確認できた．

5. おわりに

深層格選好の傾向が類似した単語をクラスタリングするために，表層情報の類似度に基づいて閾値 0.8 でクラスタリングした結果，動詞の平均 DS が分類語彙表の 0.80 から 0.90，名詞が 0.88 から 0.91 となり，分類語彙表よりも高い値を示した．また，作成されたクラスタを用いて深層格の推測実験を行った結果，精度が 61.6%から 74.9%に向上し，分類語彙表を用いた場合よりも有効であったことを確認した．

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 情報技術レターズ, vol.2, pp.91-92 (2003).

[2] 渋木英潔, 荒木健治, 桃内佳雄, 栃内香次: 誤りを含むタグ付きデータを用いた深層格の自動推測手法, 言語処理学会第 10 回年次大会発表論文集, pp.568-571 (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] S.A.Caraballo: Automatic construction of a hypernym-labeled noun hierarchy from text, Proc. of 37th Annual Meeting of the ACL, pp.120–126 (1999).

[3] 本稿で扱う深層格は，大石ら [3] と同じ，agent, object, cause, material, source, goal, place, purpose, basis, beneficiary, quantity の 11 種類である．また，意味フレームに現れない文字列を助詞とし，「を」,「に」,「は」,「が」,「で」,「から」,「と」,「も」,「へ」,「では」,「には」,「の」,「にも」,「でも」,「へと」,「まで」,「に対して」,「側は」,「のために」,「たちは」,「氏は」,「からも」の 22 種類とした．従って，表層情報による動詞ベクトルは $22 \times 22 = 484$ 次元，名詞ベクトルは位置情報を加えた $22 \times 22 \times 2 = 968$ 次元で表現され，深層格情報による動詞ベクトルは，語順を考慮しないため，$\sum_1^{11} = 66$ 次元，名詞ベクトルは 11 次元となる．

[4] 本手法は助詞パターンを利用するため，複数の助詞が含まれていることが条件であり，条件を満たす中では 2 名詞の事例が比較的豊富であったため，2 名詞の事例とした．