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

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　‡ 北海道大学大学院工学研究科
* 北海学園大学大学院工学研究科　◇ 北海学園大学ハイテク・リサーチ・センター

1　まえがき

深層格は，機械翻訳や対話処理などの知識表現に利用されており，深層格を自動的に推測するための手法が提案されている [1, 2, 3, 4]．しかしながら，これらの手法は，文単位で深層格タグが付与されたデータを学習に利用しており，新規な語彙の出現に伴う知識の拡充を行う際には，タグ付きの学習データを用意する必要がある．このような背景から，我々は，学習コストの軽減を目的とし，知識の拡充の際にタグなしのデータから学習する手法の開発に取り組んできた [5, 6]．

我々の手法は，学習の引き金となる，概念と深層格との選好指標の抽出にのみ深層格タグが付与された文を用い，その後の語彙と対応付けられた規則の獲得には深層格タグが付与されていない文を用いる．先行研究 [5, 6] では，精度を保持したまま手法の適用範囲を拡大する点に課題を残していた．本稿では，この問題に対して規則の語彙化と条件の詳細化による改善を行い，手法の全体像および改善箇所について記述し，実験結果について考察する．

2　全体の流れ

図 1 に，本手法の全体的な構成を示す．全ての学習データおよび解析データは構文解析済みとする．まず，深層格のタグが付けられた学習データ A から，概念と深層格との選好指標となる言語普遍知識を抽出する．次に，抽出された言語普遍知識を用いて，学習データ B の深層格を暫定的に推測し，タグとして付与する．その結果から，言語固有の表層表現と深層格とを対応付ける言語固有規則を獲得する．その後，言語固有規則を用いて，文の深層格の推測を行う．

このとき，言語普遍知識を用いて推測された深層格は必ずしも正解であるとは限らず，学習データ B は誤りを含むタグ付きデータとみなすことができる．したがって，誤りの影響を抑制するように言語固有規則を獲得することが望まれる．

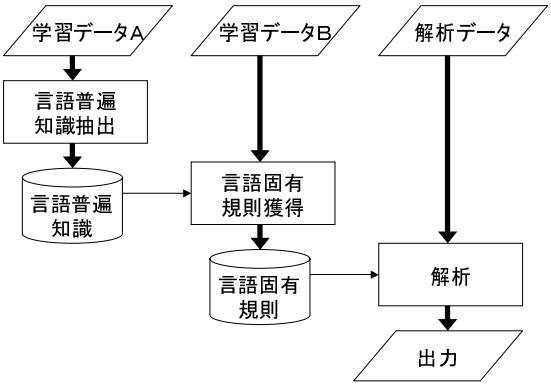

図 1: 全体の構成

言語普遍知識と言語固有知識を区別する背景には，将来的に多言語を扱う環境において，言語普遍知識を再利用したいという意図が存在している[1]．

以上の考えから，本手法では，語彙ではなく概念と対応付けられた以下の情報を言語普遍知識として定義する．

- ある概念が名詞として任意の動詞に下位範疇化される際に，その概念がなる深層格の選好確率
- ある概念が動詞として任意の名詞を下位範疇化する際に，その名詞の概念がなる深層格の選好確率

本手法では，概念として分類語彙表 [7] の上位 5 桁の数字を用い，学習データ A から以下の式に従って選好確率の計算を行った．

$$P(c,d) = \frac{F(c,d)}{\sum_{i \in D} F(c,i)} \tag{1}$$

$F(c,d)$ は概念 c が意味フレーム中で，名詞として深層格 d となる，または，d となる名詞概念を動詞としてとる回数である．D は深層格の集合であり，大石ら [1]

[1] 言語普遍知識に含まれる情報が真に言語普遍であるかという問題や，特定の言語から言語普遍知識の抽出が可能であるかという問題に関しては，現段階ではそこまで判断できる水準に達していないため，本稿では扱っていない．

表 2: 言語普遍知識の例

概念	品詞	agent	object	cause	material	source	goal	place	purpose	···
14170	名詞	0%	82%	0%	0%	0%	9%	0%	9%	···
14170	動詞	0%	33%	0%	0%	0%	0%	33%	0%	···

表 1: 深層格の一覧

agent	動作を引き起こす主体
object	動作・変化の影響を受ける対象
cause	原因
material	材料または構成要素
source	起点
goal	終点
place	事象の成立する場所
purpose	目的
basis	比較の基準
beneficiary	受益者
quantity	動作・変化の量

と同様に表 1 に示す 11 種類とした．言語普遍知識の例を表 2 に示す．表の一行目は，概念 14170 が名詞として用いられる場合に，object となる確率が 82%，goal となる確率が 9%，purpose となる確率が 9%であることを示している．同様に二行目は，概念 14170 が動詞として用いられる場合に，object となる名詞をとる確率が 33%，place となる名詞をとる確率が 33%であることを示している．

言語固有規則は，語彙や助詞などの言語に固有の表層情報を深層格と対応付ける規則である．本報告での改善点は，言語固有規則の形式と，それに伴う獲得および解析部分の変更であり，詳細は次章で述べる．

3 手法の改善

先行研究 [5, 6] では，表層情報として助詞のみを扱っており，以下の手順により言語固有知識を獲得していた．まず，言語普遍知識を用いて学習データ B の深層格を推測する[2]．この段階で推測された深層格を暫定深層格と呼ぶ．学習データ B の助詞ごとに共起する暫定深層格の頻度を計算し，最も共起頻度の高い暫定深層格と対応付ける規則を獲得する．

[2] この部分の説明は，先行研究 [5, 6] からの変更がないため，紙面の都合により省略する．

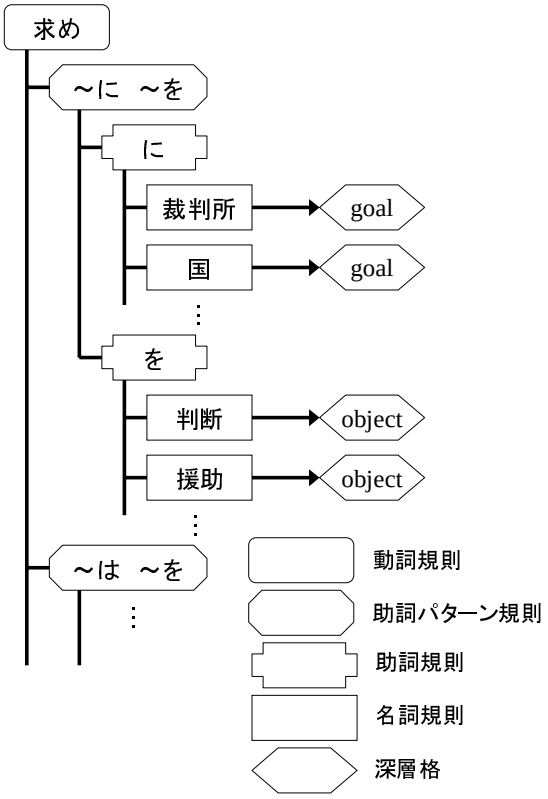

図 2: 言語固有規則の例

しかしながら，同一の概念と深層格の組であっても，どの表層格が対応するかに関しては動詞や名詞の語彙に影響を受けるところが大きい．また，格助詞以外の助詞に関しては，同一文内の他の助詞を考慮することで深層格が決定されることが多い．これらの問題に対処するため，規則の語彙化を行うことと，文内にどのような助詞が出現しているかを表す助詞パターンの情報を用いることとした[3]．本稿では，言語固有規則を，動詞規則，助詞パターン規則，助詞規則，名詞規則の四つの規則集合とし，各規則は図 2 で示すような階層構造を成している．

全ての語彙を規則化すると負担が大きいため，出現頻度の高い順に 1000 の動詞と 40 の助詞に限って規則化を行った．また，動詞，助詞パターン，助詞，名詞

[3] 本稿の助詞パターンは，文献 [4] と同じく文中での助詞の出現順序を考慮していない．それゆえ,「太郎は次郎を見た」と「次郎を太郎は見た」は同じ助詞パターンとして処理される．

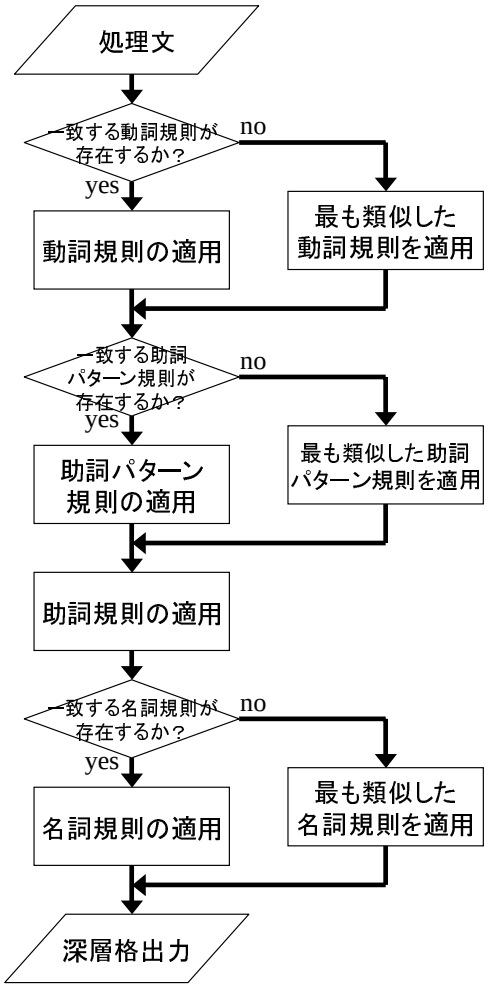

図 3: 解析の流れ

の 4 つの情報を条件としても，共起する暫定深層格が複数存在する可能性が考えられる．そのような場合には，条件を満たす事例において，最も共起頻度の高い暫定深層格を規則の深層格として獲得した．

言語固有規則を用いた解析は，図 3 に示すように，巨視的構造から微視的構造へと段階的に規則を適用することにより行われる．文全体に対して動詞規則，助詞パターン規則の順に適用した後，文内の各文節に対して助詞規則，名詞規則の順に適用する．4 つの情報を条件としたことや，高頻度の語彙に限って規則化したことにより，全ての条件と完全に一致しないことが考えられる．そのような場合には，以下の類推により最も類似した規則を適用する．

動詞規則の類推は，ベクトルで表現された 2 つの動詞の類似度に基づいて行う．上述したように，動詞規則の適用段階では，文節ではなく文単位での巨視的構造を捉えることを意図しており，文単位の構造を反映する表層情報は助詞パターンである．それゆえ，動詞のベクトルは，学習データおよび解析データにおいて共起する助詞パターンの頻度をベクトル要素として表現する．動詞 v と w の類似度は，文献 [8] と同様に以下の式に表される単位ベクトルの内積で計算する．

$$sim(\vec{v},\vec{w}) = \frac{\vec{v}\cdot\vec{w}}{|\vec{v}|\ |\vec{w}|} \tag{2}$$

助詞パターン規則の類推は，次処理として助詞規則の適用を行うため，解析対象となる文中の全ての助詞を含む助詞パターン規則を選択する必要がある．それゆえ，文中の全ての助詞を含む助詞パターン規則の中で，文中に存在しない助詞の数が最も少ない規則を，最も類似した規則として選択する．該当する規則が複数存在する場合には，頻度の高い規則を優先する．また，該当する規則が存在しない場合には，解析不能として終了する．

名詞規則の場合は，動詞規則と同様に，個々の名詞をベクトルとした場合の内積によって類推する．ただし，名詞規則の適用段階では，文節単位の微視的構造を捉えることを意図しているため，名詞ベクトルでは，助詞パターンではなく，その名詞と同じ文節内の助詞と係り先の動詞の組を要素とする．

4 実験

本稿の改善により，どの程度精度が向上するかを調査するための実験を行った．

実験で使用したデータを以下の手順で作成した．EDR コーパス [9] から表 1 の深層格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「れ」,「られ」,「れる」,「せ」,「させ」を含む事例を除外した．この中で名詞と動詞が共に分類語彙表に登録されている 84,681 事例を学習データとして用い，残りの 87,197 事例を解析データとした．本稿では，学習データ A と B の区別をせず，同一の学習データを用いた．

この学習データを用いて，改善前と改善後の手法により，それぞれ言語固有規則を獲得し，解析データの深層格を推測した．評価基準として被覆率 cov と精度 pre を用い，それぞれ以下の式により計算した．

$$cov = \frac{N_{output}}{N_{dependency}} \times 100. \tag{3}$$

$$pre = \frac{N_{correct}}{N_{output}} \times 100. \tag{4}$$

ここで $N_{dependency}$，N_{output}，$N_{correct}$ は，それぞれ，解析データ中の総係り受け数，何らかの深層格を出力した係り受け数，EDR コーパスと一致する深層格を出力した係り受け数である．

表 3: 改善による被覆率と精度の変化

	総数	出力数	（被覆率）	正解数	（精度）
改善前	87,197	71,864	(82.4%)	40,641	(56.6%)
改善後	87,197	52,214	(59.9%)	32,142	(61.6%)
学習データ B	84,681	—	(—)	54,502	(64.4%)

実験結果を表 3 に示す．言語普遍知識を用いて推測された学習データ B の暫定深層格の精度は 64.4%であった．改善前の手法では 56.6%の精度でしか推測できなかった解析データに対して，改善後の手法では 61.6%の精度で推測することができ，わずかではあるが精度が向上することを確認した．

しかしながら，規則の語彙化と 4 つの情報を条件としたことにより，被覆率が 82.4%から 59.9%に低下しており，改善前の手法が有していた広範な対象への適用能力が失われたことが示されている．この問題に対しては，規則化を行う語彙数の増大や学習データ B の追加によりある程度の対処が可能であると考えられる．また，規則獲得に用いる学習データ B には深層格タグを付与する必要がないため，追加による学習コストを従来手法よりも軽減できると考えられる．

5 まとめ

本稿では，学習コストの軽減を目的とし，知識の拡充の際に深層格タグの付与されていないデータから規則を学習する，深層格の自動推測手法の改善について報告した．我々の手法は，学習の際に，概念と深層格との選好確率を基にタグなしデータの深層格を推測し，その結果をタグとして利用することで規則を獲得する．このとき推測された結果には誤りが含まれており，誤りの影響を抑制するように規則を獲得することが課題であった．この問題に対して，規則の語彙化と，動詞，助詞パターン，助詞，名詞の 4 つの情報を条件とすることによる改善を行った．64.4%の精度でタグが付与された学習データから獲得された規則を用いて実験を行った結果，解析データの深層格の推測精度が 56.6%から 61.6%に向上したことを確認した．

本稿の改善により，誤りを含んだ学習データと近い精度で解析データの推測を行えることを確認したが，学習データの精度を超えるには至らなかった．今後は，規則化の際に誤りの影響を抑制する現在のアプローチに加えて，言語普遍知識を用いた推測手法の改善を行い，学習データに対するタグ付けの精度を向上させることで，更なる精度の向上を行いたいと考えている．

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] 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).

[2] Gildea, D. and Jurafsky, D.: Automatic Labeling of Semantic Roles, *Proc. 38th ACL* (2000).

[3] 峨家正樹, 荒木健治, 栃内香次: 帰納的学習を用いた言語に依存しない意味解析手法の評価, 情報処理学会研究報告, NL-146-12, (2001).

[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] 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).

[6] 渋木英潔, 荒木健治, 栃内香次: 一文一格の原理と深層格選好に基づいた日本語深層格規則の自動獲得手法, 情報処理学会研究報告, NL-156-3, (2003).

[7] 国立国語研究所: 分類語彙表, 秀英出版 (1964, 1993).

[8] Caraballo, S.: Automatic construction of a hypernym-labeled noun hierarchy from text, *Proc. 37th ACL* (1999)

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