コーパスサイズの拡大および用例の汎化による格フレームのカバレッジの改善

笹野 遼平　　河原大輔　　黒橋 禎夫
東京大学大学院情報理工学系研究科　情報通信研究機構　京都大学大学院情報学研究科
ryohei@nlp.kuee.kyoto-u.ac.jp　dk@nict.go.jp　kuro@i.kyoto-u.ac.jp

1　はじめに

用言とそれがとる格要素の関係を記述した格フレームは、高精度な格解析や省略解析を実現する上で必要となる知識であり、近年、大規模な格フレームをコーパスから自動構築することが可能となっている [3]。しかしながら、格フレーム構築に用いていないコーパスを使い、格要素がその係り先用言の格フレームの用例にどの程度含まれているかを調べると、ウェブテキスト約 5 億文から自動構築した格フレームであっても、そのカバレッジは 60%程度に過ぎない [4]。このため、高精度な格解析や省略解析を行うためには、よりカバレッジの大きい格フレームが必要になると考えられる。

格フレームのカバレッジを改善する方法としては、格フレーム構築に用いるコーパスサイズを拡大することにより格フレームの用例を増やす方法や、格フレームの用例を汎化することにより間接的に格フレームのカバレッジを拡大する方法などがある。本稿ではまず、より大規模なコーパスを利用、および、格フレームの用例の汎化により、どのくらい格フレームのカバレッジが改善されるかを調査する。

続いて、構築した格フレームを構文・格解析の統合的確率モデルに適用することにより、コーパスサイズの拡大、および、用例の汎化により構築されたカバレッジの大きな格フレームの有用性を示す。

2　格フレーム構築とコーパスサイズ

格フレームは、ウェブから収集したコーパスを用いて、河原ら [3] の手法により自動構築を行う。格フレーム構築手法の概要は以下の通りである。

1. KNP を用いてコーパスを構文解析し、構文的曖昧性のない述語項構造を抽出する。
2. 抽出した述語項構造を用言とその直前の格要素のペアごとにまとめ、α 回以上出現したペアを用例パターンとして収集する。

表 1: 「目指す」の格フレーム

用言:目指す-動詞-1	
格	用例 (素性)
ガ格	選手:6 監督:6 マリオ:6 私:4 彼女:4 韓国:2 市:1 . . . <CT:人>:0.368 <NE:PERSON>:0.120 <CT:組織・団体>:0.057 <NE:LOCATION>:0.048 . . .
ヲ格*	向上:11915 発展:3190 飛躍:749 強化:551 開始:520 改善:499 . . . <CT:抽象物>:0.985 . . .
⋮	
用言:目指す-動詞-8	
格	用例 (素性)
ガ格	私:63 我々:20 彼:11 あなた:10 僕:7 学生:4 里奈:2 . . . <CT:人>:0.665 <NE:PERSON>:0.052 . . .
デ格	大学:5 学校:4 東京:4 分野:4 ロンドン:3 海外:2 日本:2 会社:2 . . . <NE:LOCATION>:0.364 <CT:場所-施設>:0.132 <CT:組織・団体>:0.066 . . .
ヲ格*	教師:519 デザイナー:400 女優:257 医師:238 歌手:220 女:185 税理士:156 . . . <CT:人>:0.896 . . .
⋮	*は直前格であることを示す

3. 用言ごとに、用例パターンの出現数上位 β 個を抽出し、クラスタリングを行う。
4. クラスタリングに用いられなかった用例パターンを、最も類似しているクラスタに振り分ける。

最終的に出来上がった各クラスタが格フレームである。表 1 に格フレームの例を示す。

格フレームの格スロットの用例は、構築に用いたコーパス中に出現した格要素から生成されるため、基本的にコーパスサイズが拡大するにつれカバレッジは増大する。そこで本研究では、様々な規模のコーパスを用いて格フレームの構築を行うことにより、コーパスサイズがどの程度の格フレームのカバレッジに影響するか、どのくらいのコーパスサイズがあれば十分であるかの調査を行う。

3　格フレームの用例の汎化

格フレームの同一の格スロットの用例には、類似した性質を持つ表現が集まる。例えば表 1 に示した格フレーム「目指す-動詞-8」の場合、ガ格には “人”、ヲ格には “職業”、ニ格には “場所” を意味する用例が集

まっている。格フレームを用いて格解析や省略解析などの解析を行う場合、文章中に出現した表現を適切な格スロットに対応付ける必要があるが、「目指す」という用言の近くに「作家」などといった “職業” を表す表現が出現した場合は、この格フレームのヲ格に対応付けるべきであると考えられる。しかしながら、コーパスを増やすことにより用例の数を増やすことは可能であるものの、固有表現を含むすべての用例を集めることは現実的ではなく、対応付けるべき表現が必ずしも格スロットの用例に含まれているとは限らない。

そこで本研究では、格スロットに集まった用例を汎化し、汎化した用例をどのくらい含むかという汎化情報を格スロットに付与することにより、このような対応付けを可能にする。汎化情報としては以下の 2 種類を考える。

固有表現情報 格フレーム構築に用いるコーパスに対して事前に固有表現抽出を行い、固有表現抽出済みのコーパスを用いて格フレームを構築することにより、固有表現情報を格フレームに付与する。

抽出する固有表現は、IREX[1] で定義された 8 種類の固有表現とし、CRL(郵政省通信総合研究所) 固有表現データを学習データとして、CRF を用いて学習した固有表現抽出器により固有表現抽出を行う。この固有表現抽出器のウェブテキストに対する精度 (F 値) はおよそ 0.7 である。

固有表現の主辞であると解析された用例を対応する固有表現に汎化し、各固有表現の割合を格スロットに以下の形式で記述する。

<NE:(固有表現名)>:(用例に占める割合)

カテゴリ情報 形態素解析器 JUMAN 6.0[5] では、辞書に登録されているすべての普通名詞、サ変名詞に、約 20 種のカテゴリが付与されている。例えば、「選手」には “人”、「市」には “組織・団体”、および、“場所-その他” というカテゴリが付与されている。

本研究では、形態素解析器として JUMAN 6.0 を使用することにより、普通名詞、サ変名詞にカテゴリを付与し、これらの情報を汎化情報として、以下のような形式で格スロットに記述する。

<CT:(カテゴリ名)>:(用例に占める割合)

表 1 に示した格フレームには汎化情報がすでに付与されており、例えば「目指す-動詞-8」のガ格には、固有表現情報として “<NE:PERSON>:0.052”、カテゴリ情報として “<CT:人>:0.665” がそれぞれ付与されている。

4 汎化情報を用いた構文・格解析

4.1 構文・格解析の統合的確率モデル

河原ら [4] は、格フレームによる語彙的な選好を利用した構文・格解析の確率モデルを提案している。このモデルでは、入力文がとりうるすべての構文構造に対して確率的格解析を行い、もっとも確率値の高い格解析結果をもつ構文構造を出力する。

このモデルにおいて格フレームは、用例生成確率 $P(n_j|CF_l, A(s_j) = 1, s_j)$ を推定するのに使用される。用例生成確率とは、ある格フレーム (CF_l) のある格スロット (s_j) に何らかの入力側格要素が対応付けられている ($A(s_j) = 1$) 場合に、その入力側格要素が n_j となる確率のことであり、格要素生成確率を計算する際に必要となる。

河原らはこのモデルを用いることにより、従来手法より高い構文解析の精度を得ている。しかしこの手法には、用例生成確率を計算する際、入力文に出現した格要素が対応する格フレームの格スロットの用例に含まれていない場合に、用例生成確率を適切に推定できないという問題がある。河原らはこのような場合、人手で設定した確率 γ を用例生成確率として与えているが、よりカバレッジの大きな格フレーム、汎化情報の付与された格フレームを用いることによりこの問題は改善できると考えられる。

4.2 汎化情報を用いた用例生成確率の推定

入力文に出現した格要素が対応する格フレームの格スロットの用例に含まれていない場合でも、対応する格スロットに類似した用例が多く含まれている場合は、その用例生成確率はある程度高くなるべきだと考えられる。そこで本稿では、格要素が用例に含まれず、かつ、その格要素が主辞となっている固有表現、または、その格要素が属するカテゴリの情報が対応する格スロットに記述されている場合に、用例生成確率 $P(n_j|CF_l, A(s_j) = 1, s_j)$ を以下の式により近似する手法を提案する。

$$P(n_j|CF_l, A(s_j) = 1, s_j) = P(n_j|GE) \times P(GE|CF_l, A(s_j) = 1, s_j)$$

ただし、GE は汎化された用例 (Generalized Example) を表わし、実際には対応する固有表現またはカテゴリとなる。この式を用いることにより、例えば表 1 に示した「目指す-動詞-8」のヲ格に何らかの入力格要素が対応付けられている場合にそれが「作家」となる確率は、用例に「作家」が含まれていない場合でも、

$P($ 作家 |CT:人 $) \times P($CT:人 | 目指す-動詞-8, ヲ格 $)$

表 2: 構築した格フレームの統計情報

コーパスサイズ (文)	160 万	630 万	2500 万	1 億	4 億	16 億
用言数	2684	6751	14939	30194	57756	106374
(内訳) 動詞	2208	5420	11433	21404	38099	65681
形容詞	179	367	685	1343	2406	4244
名詞+判定詞	297	964	2821	7447	17251	36449
用言あたりの平均格フレーム数	3.2	4.1	5.0	5.6	6.2	6.6
格フレームあたりの格スロットの平均数	3.6	4.0	4.5	4.9	5.3	5.6
格スロットあたりの平均用例数	13.7	22.1	39.1	71.0	144.6	279.6
格スロットあたりの平均異なり用例数	2.3	3.7	6.2	9.5	16.1	24.9
格スロットあたりの汎化情報数	1.0	1.2	1.6	2.0	2.4	2.8

という式により計算できるようになる。n_j が固有表現となる場合も同様である。

5 実験

5.1 格フレームのカバレッジ

ウェブテキスト約 160 万文、630 万文、2,500 万文、1 億文、4 億文、16 億文からそれぞれ、用例の汎化情報を含む格フレームの構築を行った。いずれの場合も用例パターンとして収集する閾値 α は 5、クラスタリングに使用する用例パターンの数 β は 20 とした。構築された格フレームの統計情報を表 2 に示す。

用言数や平均格フレーム数、格の平均数などいずれの数値もコーパスサイズの増加とともに大きくなっている。先行研究と比べて平均格フレーム数が少ないのは、クラスタリングに使用する用例パターンの数を 20 に制限したため、20 を超える数の格フレームを持つ用言が存在しないためである。

続いて、使用するコーパスサイズ、および、用例の汎化による格フレームのカバレッジの変化を調べるため、人手により用言とその格要素の関係が付与された京都テキストコーパス [2] を用いて、用言の格要素のうち、対応する格フレームが構築されているものの割合を調べた。汎化情報を用いない場合、用いる場合、それぞれにおける、対応する格フレームが構築されていると判断する基準は以下の通りである。

汎化情報なし 用言の格要素が対応する格フレームの格スロットの用例に含まれている。

汎化情報あり 上記の場合に加え、用言の格要素が主辞となっている固有表現、または、属するカテゴリ情報が、対応する格スロットに付与されている。

格要素と用言が直接係り受け関係にある場合のカバレッジを図 1 に、係り受け関係にない、すなわち、格要素が省略されている場合のカバレッジを図 2 に示す。

格要素との係り受け関係の有無に関わらず、コーパスサイズの拡大とともに格フレームのカバレッジは増加している。また、汎化情報を用いた対応をとることにより格フレームのカバレッジは大きく改善している。

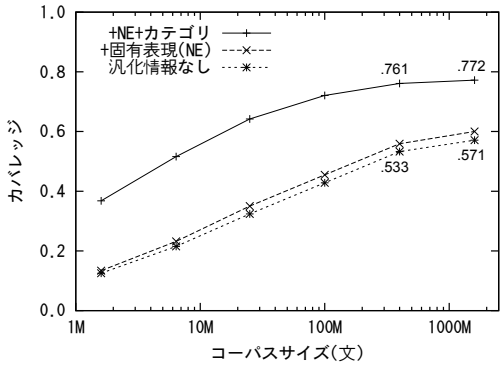

図 1: 格フレームのカバレッジ (係り受けあり)

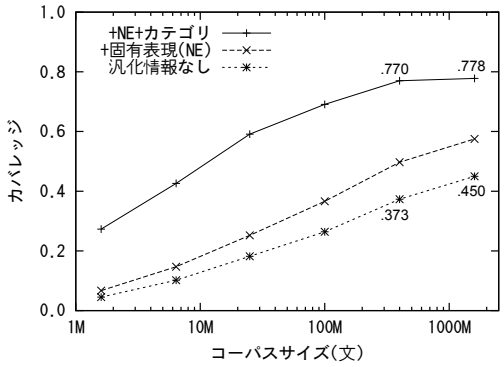

図 2: 格フレームのカバレッジ (係り受けなし)

汎化情報を用いた場合の 16 億文から構築した格フレームのカバレッジは、係り受け関係がある場合は 77.2%、係り受け関係がない場合は 77.8%に達した。ただし、いずれの場合も、コーパスサイズが 4 億文の場合でもそれぞれ 76.1%、77.0%のカバレッジとなっており、これ以上コーパスサイズを増やしてもカバレッジの改善はあまり見込めないと考えられる。

一方、汎化情報を用いない場合のカバレッジは、コーパスサイズが 4 億文から 16 億文に増えることにより、それぞれ 3.8%、7.7%増加している。このため、汎化情報を用いた場合との差は小さくなってきており、コーパスサイズの増加とともに、汎化情報と用いることによるカバレッジの改善効果は小さくなってくると考え

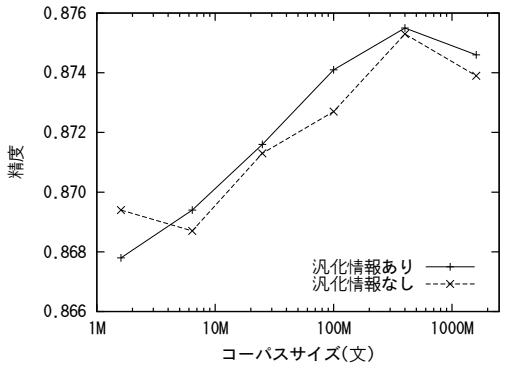

図 3: 構文解析の精度

られる。ただし、汎化情報として固有表現のみを使用した場合との差は小さくなっておらず、固有表現情報を用いることによるカバレッジの改善は、コーパスサイズが十分に大きくなっても有効であると考えられる。このことは、固有表現の数が膨大であり、本質的に網羅できないためであると考えられる。

5.2 構文解析実験

コーパスサイズの拡大、および、用例の汎化による格フレームのカバレッジの増大の有用性を確認するため、構築した格フレームを用いて、河原らの統合的確率モデルに基づく構文解析実験を行った。

実験には、京都テキストコーパスと同じ基準で、4409 個の係り受けのタグが付与されたウェブテキスト 759 文を用いた。また、統合的確率モデルを生成するのに必要となる、格フレーム以外のリソースから計算されるパラメータは、ウェブテキスト約 600 万文をシソーラスに基づく類似度を用いた格解析を行うことによって得られた格解析済みデータと、京都テキストコーパスを用いて計算した。これらのリソースは、格フレーム構築に用いたコーパスサイズに関わらず、すべて同一のものを用いている。構文解析実験の結果を図 3 に示す。

全体的な傾向としては、コーパスサイズの増加とともに構文解析の精度も向上しており、基本的には、格フレーム構築に用いるコーパスを増やすことにより構文解析の精度も上昇すると言える。しかし、コーパスサイズを 4 億文から 16 億文に増やすと精度は僅かではあるが低下している。このことは、コーパスサイズの増加により、用言数、平均異なり用例数がそれぞれ約 2 倍、1.5 倍となっているのに対し、格フレームのカバレッジの増加率は 10%に満たないことから、コーパスを増やした結果、適当でない用例の数が増えたために引き起こされたと考えられる。コーパスを増やすことにより適当でない用例の数が増えた原因としては、コーパスサイズに関わらず用例パターンとして収集する閾値 α を同じ値にしており、また、閾値 α を満たす用例パターンに含まれる用例をすべて収集しているためであると考えられ、用例の絞り込み適切に行うことによりこの問題は解消できると考えられる。

また、汎化情報を用いることにより僅かではあるが精度が向上する傾向が確認できた。以下に固有表現情報、カテゴリ情報が有効に働き、構文解析結果が改善された例を 1 つずつ示す。

(1) ...大阪 で料理人を**目指す**$_{\bigcirc}$ 姿を**描く**$_{\times}$ 物語。
(2) ... 時刻 表 を**わかり**$_{\times}$ やすく、使いやすく**提供**$_{\bigcirc}$ する「いわて時刻表」。

(1) の文では、固有表現情報を使用しない場合、「目指す」のデ格の用例に「大阪」が存在しないため、「大阪」の係り先が「描く」であると誤って解析されていたが、「目指す」のある格フレームのデ格に占める “LOCATION” の割合が 0.364 であり、“LOCATION” に占める「大阪」の割合が 0.0081 であることから、格スロットに記された固有表現情報を使用することにより、「目指す」のデ格の用例が「大阪」となる確率は比較的大きいと計算されるようになり、「大阪」が「目指す」に係ると正しく解析できるようになった。(2) の文も同様にカテゴリ情報を使用することにより、「表」の係り先が「わかり」ではなく「提供」であると正しく解析できるようになった。

6 おわりに

本稿では、使用するコーパスサイズを約 160 万文から 16 億文まで 6 段階に変化させた上で、汎化情報を含んだ格フレームを構築し、それらのカバレッジの調査を行った。また、構築した格フレームを構文・格解析の統合的確率モデルに適用し、カバレッジの大きな格フレームの有用性を示した。今後の課題としては、適切な用例の絞り込みを行い、適当でない用例を含まない格フレームを構築することが考えられる。

参考文献

[1] IREX 実行委員会（編）. IREX ワークショップ予稿集, 1999.
[2] Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. Construction of a Japanese Relevance-tagged Corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pp. 2008–2013, 2002.
[3] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた web からの大規模格フレーム構築. 情報処理学会 自然言語処理研究会 2006-NL-171, pp. 67–73, 2006.
[4] 河原大輔, 黒橋禎夫. 自動構築した大規模格フレームに基づく構文・格解析の統合的モデル. 自然言語処理, Vol. 14, No. 3, 2007.
[5] 黒橋禎夫, 河原大輔. 日本語形態素解析システム JUMAN version 6.0 使用説明書. 京都大学大学院 情報学研究科, 9 2007.