大規模コーパスからの格フレーム辞書構築とそれを用いた格解析

河原大輔
京都大学大学院情報学研究科

鍛冶伸裕
京都大学工学部

黒橋禎夫
京都大学大学院情報学研究科

{kawahara, kaji, kuro}@pine.kuee.kyoto-u.ac.jp

1 はじめに

自然言語処理における文レベルの解析では、これまで構文解析の問題が主に扱われていた。しかし、日本語には語順の入れ替わり、格要素の省略、表層格の非表示などの問題があり、構文解析を行っただけでは文の解析として十分とはいえない。これらの問題を解決するためには、格フレーム辞書に基づく格解析が必要である。

ひとつの方法は、人手によって格フレーム辞書を構築することである [2, 7]。しかし、その場合には非常に大きなコストがかかり、カバレージの大きな辞書を作成、保守することは簡単ではない。

そこで、格フレーム辞書をコーパスから自動学習する方法が考えられる [3]。格フレームとは用言と体言の組み合わせを記述するものであるから、それを学習するためには膨大なデータが必要となるが、現存するタグ付きコーパスはこのような目的からは量的に不十分である。本論文では、格フレーム辞書をタグ情報が付与されていない大規模コーパスから自動的に構築する手法と、得られた辞書を用いて格解析を行う手法を提案する。

2 格フレーム辞書の構築

前節で述べたように、格フレーム辞書の人手による作成はコスト的に問題がある。また、これまでの人手による辞書ではガ格、ヲ格、ニ格などの基本的な格は扱っているが、「によって」、「として」などの格と同じ振る舞いをする複合辞を取り扱っているものは少ない。

そこで、辞書をコーパスから自動学習することを考える。タグ付きコーパスの場合には、構文情報が明示されているという利点があるが、一方でカバレージに問題がある。例えば、約 4 万文の京都大学テキストコーパス (毎日新聞 2,900 記事) でも非常に一般的な用言で頻度が 10 回以下のものが多数ある (例えば、「寂しい」の出現回数は 6 回、「助ける」は 10 回)。そのような少数のデータから用言の振る舞いを学習することは難しい。

そこで、本稿ではタグの付与されていないコーパス、すなわち一般のテキストからの学習方法を提案する。この方法では膨大な量の用例を学習対象とすることができる。本稿で提案する格フレーム辞書の構築手順は以下のとおりである。

1. (格フレーム辞書を用いない) 頑健で高精度なパーサである KNP[5] を用いる。KNP は、新聞記事に対して約 90% の精度で係り受けの解析を行うことができる。
2. このパーサを用いて大量のコーパスを解析する。我々の実験では毎日新聞 360 万文を対象にした。
3. 解析済みの文から、ある程度信頼できる用言・体言間の関係を取り出し、格フレーム辞書としてまとめる。

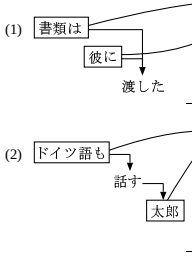
以下では上記 3 の格フレームの収集について詳しく述べる。

2.1 格フレームの自動収集の問題点

コーパスを解析した結果から、格要素と用言の組み合わせを収集し、それらを用言ごとにまとめて整理することによって、格フレームを作ることができる。しかし、そこでは次のような問題がある。

意味の多義性
: 同じ表記の用言でも、意味によって格パターンが変わる。

構文の曖昧性
: パーサによって出力された解析結果には、誤りが含まれる。



用言	格	用例
手伝う	ガ	夫, 人, 子供, 職員, 私, 容疑者, 教諭, 日本, 自分, ...
	ヲ	仕事, 作業, 家事, 家業, 店, 活動, 農作業, 選挙, ...
	ニ	息子, 大使, 友達, 会員, お礼, 人, 夫, 事務所, ...
	デ	ボランティア, 事件, 実家, 事務所, 報酬, 本部, ...
咲く	ガ	花, 桜, サクラ, バラ, 梅, アジサイ, ツツジ, ...
	ニ	話, 庭, 野, 談議, 夜空, 初夏, 斜面, 思い出話, ...
	デ	各地, 公園, 幡, 東京, 周囲, 日本, 奈良, 全国, ...
転がる	ガ	死体, 空き缶, 玉, 体, 岩, 球, 骨, ボール, 溶岩, ...
	ヲ	斜面, 上, 坂道, 坂, 床, 道, 中間, 地面, 道路, ...
	ニ	床, 地面, 前, 上, 周り, その辺, 土俵, 足元, 隅, ...
渡す	ガ	被告, 容疑者, 社長, 幹部, 氏, 私, 女性, 社員, ...
	ヲ	現金, 金, 書, カネ, わいろ, 政権, メモ, 引導, ...
	ニ	容疑者, 被告, 総合屋, 側, 者, 人, 幹部, 知事, ...
	カラ	中, 財布, 窓, 震災, 金, 役員, 資金, 氏, バッグ, ...
	デ	名目, 事件, 趣旨, 事務所, 室, 市内, 目的, 小切手, ...
	トシテ	報酬, 謝礼, 資金, 献金, 対策, 費, 見舞い, 担保, ...
	ニヨッテ	判決, 起訴状, 調べ, 関係者, 証言, 供述, 訴状, ...
話す	ガ	監督, 被告, 女性, 私, 氏, 人, 容疑者, 者, 教授, ...
	ヲ	こと, 理由, 体験, 日本語, 感想, 語, 英語, 動機, ...
	ニ	取材, 人, 関係者, 友人, 調べ, 周辺, 家族, 質問, ...
	ト	こと, 人, 問題, 課題, 資料, 気持ち, 子供, 友人, ...
	カラ	経験, 立場, 方, 結果, 側, 普段, 私, 経緯, 経験, ...
	デ	電話, 会見, 表情, 口調, 声, 様子, 課, 調べ, 日本語, ...
	ニツイテ	こと, 理由, 結果, 問題, 点, 事件, 状況, 判決, ...
	ニタイシテ	これ, 取材, 調べ, 聴取, 調査, こと, 取り調べ, ...

図 1: 格の対応付けと格フレームの例

2.2 意味の多義性の扱い

同じ表記の用言でも、用言の意味によって格パターンが変わることがある。例えば

病気になる: ～が病気になる
親密になる: ～が～と親密になる

この例では、「親密になる」はト格をとるが、「病気になる」はト格をとらないというように格パターンが変化している。しかし、この現象は頻度が高く、意味の多義性の多いある種の限られた用言にしか起こらない。

このような用言は IPAL[7] に登録されており、また重要な格要素は用言の直前にくることが多いので、IPAL に登録されている動詞、形容詞について、直前の名詞句とのペアでひとまとまりの用言として扱い、それらは別々の格フレームとして集めることにする。その他の用言については、格パターンはひとつであると仮定する。

2.3 構文の曖昧性の扱い

構文解析結果には誤りが含まれているので、解析の精度が低い係り受けは捨てて、ある程度確信度が高い係り受けを格フレームの収集に用いた。具体的には、次のような係り受けを捨てることにした。

- 提題助詞をもつ格要素と用言の連体修飾先 (これらは表層格が明示されていないため扱わない)
- 「～で」の形をしていて、判定詞かデ格の区別がつかないもの
- 用言が受身または使役の場合

さらに、集まった結果において、各用言ごとに頻度が少ない格は除いた。これは、ひとつには構文解析結果の誤りへの対策であり、また頻度の少ない格はその用言と関係が希薄であると考えられるからである。頻度の閾値は、現在のところ経験的に $3\sqrt{mf}$ と定めている。ただし、mf はその用言において最も多く出現した格の延べ用例数である。

2.4 格フレーム辞書の構築結果

毎日新聞約 7 年分の 360 万文から実際に格フレーム辞書を構築した。格フレームの例を図 1 の右側に示す。

82,000 個の用言について格フレームが集まった。用言あたりの格の平均数は 2.5 個、ひとつの格あたりの平均異なり用例数は 16.0 個であった。

図 1 に示すように、自動構築の結果の多くは人間が見ても妥当なものになっている。しかし、格フレーム辞書の静的な評価は難しいので、次節のようにこれを解析に用いることによって評価した。

3 格フレーム辞書に基づく解析

自動的に構築した格フレーム辞書と、それを用いた格解析をパーサの中に組み込んだ。この格解析の基本的な処理は、入力文と格フレームとの対応付け (格の解釈) である。パーサは、いろいろな制約のもとでとりうる構造のひとつひとつについて格の対応付けを行って、その構造の妥当性をスコアにして、最もスコアの高い構造を出力する (基本的には [1] のアルゴリズム)。

まず格の対応付けについて述べ、次に格解析を組み込んだ構文解析について説明する。

3.1 入力文と格フレームの対応付け (格の解釈)

入力文中のある格要素と格フレームとの対応付けを行うとき、格要素の表層格が明示されているかどうかで次のように処理が異なる。ここで、表層格が明示されているとは、格要素に格助詞が付いている場合であり、表層格が明示されていないとは、格要素に提題助

詞だけがついている場合、または被連体修飾詞の場合である。

表層格明示

格要素を、その表層格の格スロットへ対応付ける。

表層格非明示

ガ格、ヲ格、ニ格[1] のうち、格要素と最も用例が類似度している格スロットへ対応付ける。

ただし、同一の格スロットに複数の格要素を対応付けることは行わない。また、いくら大規模なコーパスから用例を集めたといっても、単語そのもののマッチングでは一致する場合が多くないので、シソーラスを用いた般化を行う。入力文の格要素と格スロットの用例の類似度は、分類語彙表 [4] の一致レベルに応じて次のように求める[2]。

レベル	0	1	2	3	4	5	6	一致
類似度	0	0	5	7	8	9	10	11

図 1 の左側に、対応付けの例を挙げる。(1) では、まず「渡す」のニ格は「彼に」で埋まり、「書類は」はガ格、ヲ格の用例と類似度を計算し、より類似しているヲ格に対応付けられる。(2) では、「ドイツ語」、「太郎」の両方の格がわからないが、格フレームとの類似度をとることにより、ドイツ語がヲ格、太郎がガ格に対応付けられる。

3.2 格解析に基づく構文解析

格解析を組み込んだ構文解析とは、格解析を行うと同時に、格解析を利用して構文的曖昧性を解消する処理である。

日本語文の典型的な曖昧さは、

名詞 $_1$–格助詞 動詞 $_{1(連体)}$ 名詞 $_2$–格助詞 動詞 $_2$

という構文において「名詞 $_1$–格助詞」が動詞 $_1$ または 動詞 $_2$ にかかるという問題である。従来の KNP では、読点がなければ動詞 $_1$、 読点があれば動詞 $_2$ に係ると解析していた。読点がなければ近くに係ることが圧倒的に多く、読点があれば遠くへ係ることが多いからである。

これに対して格解析に基づく構文解析では、格要素が、動詞 $_1$ の格フレームにある用例と動詞 $_2$ の用例とのどちらが意味的に類似しているかを調べる。

[1]ニ格は被連体修飾詞にのみ対応付けが可能であるとする。

[2]格フレームには、ひとつの格スロットに複数の用例があるので、その中で最大の類似度とする。

表 1: 提題, 連体修飾の格の解釈の評価

	正解	誤り			
		対応付けの誤り	外の関係による誤り	ガガ構文による誤り	係り受けの誤り
提題	75	2	–	12	11
連体修飾	64	7	18	–	11

そのためには、結局全体のバランスを見る必要があるので、全ての構文の可能性を考え、それぞれについて格フレームの対応付けのスコアを計算し、スコアの総和が最大の構文を選ぶということを行う。

ただし、従来の KNP のヒューリスティックスは非常に有効なので、読点がなければ一つ目、二つ目、三つ目、…の動詞との係り受けに 0, –2, –4, ... というペナルティを与える。また、読点があれば –2, 0, –2, –4, ... というペナルティを与える。

全ての可能な構造に上記の計算をするが、並列構造は先に推定しており、強い従属節や、引用符などがある程度構造を制限するので、実際には構造が爆発してしまうことは少ない。例えば、次に示す新聞記事 2,720 文の実験でも、解析速度は 3 文/sec であり、一文の解析時間を 20 秒に制限しても、99.8% の文が解析できる。

3.3 実験と評価

京都大学テキストコーパス [6] を用いて実験を行った。このコーパスは、毎日新聞の約 4 万文に形態素情報、構文情報を付与したものである。このコーパスのうち 2,720 文 (2 日分の新聞記事全文) をテストセットとし[3]、これに対して格解析行った結果を評価した。解析結果の例を表 2 に示す。

格解析による、提題、連体修飾の格の解釈の結果を評価したものを表 1 に示す。誤りは、対応付けのスコア計算によって誤った格に対応付けられたもの、被連体修飾詞が外の関係の名詞であるもの (「〜する方針を」のような例)、ガガ構文であるとき (現在の格フレーム学習ではひとつのガ格しか学習していない)、そして係り受けが誤っているものである。対応付けのスコア計算による誤りは非常に少なく、提案する格フレーム学習の手法やそれに基づく解析が優れていることがわかる。このうち、外の関係、ガガ構文による誤

[3]このテストセットは、格フレーム辞書の構築には用いていない。

りは非常に難しい問題を含んでいるので、これらについては今後の課題である。

一方、構文解析の精度についてみると、係り先の変化のあった文節は 426 個あり、そのうち係り先が正しくなったものが 189 文節、誤りになったものが 186 文節であり、わずかであるが精度が向上した (解析精度としては 89.9% で変化がなかった)。

4 終わりに

本論文では、タグ情報が付与されていない大規模コーパスから、格フレーム辞書を自動的に構築する手法を提案した。得られた辞書を用いて実際に格解析を行った結果、提題、連体修飾の格の解釈をかなり高い精度で行うことができた。今後、対話処理や要約処理などでこの格解析システムを利用していく予定である。

参考文献

[1] S. Kurohashi and M. Nagao. A method of case structure analysis for japanese sentences based on examples in case frame dictionary. In *IEICE Transactions on Information and Systems*, Vol. E77-D No.2, 1994.

[2] NTT コミュニケーション科学研究所. 日本語語彙大系. 岩波書店, 1997.

[3] 宇津呂武仁, 宮田高志, 松本裕治. 最大エントロピー法による下位範疇化の確率モデル学習および統語的曖昧性解消による評価. 情報処理学会 自然言語処理研究会 97-NL-119, pp. 69–76, 1997.

[4] 国立国語研究所. 分類語彙表. 秀英出版, 1964,1993.

[5] 黒橋禎夫, 長尾眞. 並列構造の検出に基づく長い日本語文の構文解析. 自然言語処理, Vol. 1, No. 1, 1994.

[6] 黒橋禎夫, 長尾眞. 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 回年次大会発表論文集, pp. 115–118, 1997.

[7] 情報処理振興事業協会技術センター. 計算機用日本語基本動詞辞書 IPAL (Basic Verbs) 説明書. 1987.

表 2: 解析結果の例

提題助詞について格の解釈が正しく行われた例

(1) 防衛庁は$_{ガ格}$ 四日までに、一九九六年度からの新たな在日米軍駐留経費負担についての米側との対処方針を固めた。

(2) マスメディアの 力も$_{ヲ格}$ 借りたい。

被連体修飾詞について格の解釈が正しく行われた例

(3) この制度から 起きる 諸問題に$_{ガ格}$ ついて国民運動が必要だ。

(4) 彼が当時、持っていた 自社株の$_{ヲ格}$ 総額は、当時の株価で二千五百万円は超えていたはずだ。

係り受けが正しくなった例

(5) キッシンジャー元国務長官は米紙 ワシントン・ポストに「直ちにＮＡＴＯを拡大せよ」と 題する$_{O}$ 論文を発表。$_{N}$

(6) ２９回目を迎えるアメリカズカップの 歴史に、新たな一ページを 刻む$_{N}$ 時が やってきた。$_{O}$

(7) ２年前の 世界選手権で、男子３、女子１の金メダルを獲得した$_{N}$ 開催国・日本がお家芸の座を 守れるか。$_{O}$

(8) 準決勝は ３回の 総当たり戦で、$_{N}$ 勝ち数の多い２チームが決勝へ 進む。$_{O}$

(9) 久山さんが 挙げる 国際人の$_{O}$ 条件だ。$_{N}$

(10) 一方、連邦情報局は昨年 多発した ロシア南部での$_{O}$ ハイジャック事件で、$_{N}$ 最終的に犯人にチェチェン領内に逃げ込まれ、じだんだを踏んだ。

係り受けが誤りになった例

(11) 制服姿で現れた両選手は同庁通信指令本部センターで、強盗事件を想定した模擬通報を受理、各警察署に 無線で 緊急配備を 指令する$_{O}$ など 緊張した$_{N}$ 表情だった。

(12) 九条署の 調べでは、居間の雨戸、ガラス窓を貫通して壁に 達した$_{N}$ 穴のほか、二階の屋根のひさし、外壁にそれぞれ一カ所ずつ計三カ所に弾痕が あった。$_{O}$

(13) 松下総裁によれば、ティートマイヤー独連銀総裁など、さくら銀行会長時代からの顔見知りも多く、唯一アジアを 代表する 日銀の$_{O}$ 新総裁を$_{N}$ 「温かく迎えてくれた」という。

(14) 燃え上がる 落日を$_{O}$ 背に、$_{N}$ 法隆寺五重塔のシルエットが浮かび上がる。

注) O の下線部が従来の解析の係り先、N の下線部が格解析を行ったときの係り先を示す。