やさしい日本語格フレームの構築による係り受け解析

角張竜晴　　山本和英
長岡技術科学大学
{kakubari, yamamoto}@jnlp.org

1 はじめに

日本語には文章中の語順の入れ替わりや格要素の省略といった問題があり、単純な係り受け解析だけでは文の構造を正しく把握できない。例えば、文1の場合、“カメラで”が“走っている”または“見た”のどちらに係るのか計算機は判断できない。

1. 私がカメラで走っている少女を見た。

計算機に文章を正しく理解させるためには、文法や単語間の関係などの知識が必要であり、それらは人間が持っている幅広い知識が重要である。それらの知識を表したものの一つに“格フレーム”がある。格フレームは、用言とその用言がとる格要素の関係を表したものである。例えば、“かける”という用言の格フレームの一つに次のようなものが考えられる。

2. {私, 父, 学生, …} が {家, 会社, 学校, …} に かける

このような格フレームを構築するには、用言の多義性を考慮しなければならない。“かける”の場合、上記の文2は“走る”と同じような意味であるが、下記の文3は“吊るす”と同じような意味として使われている。このように、“かける”の一語でも複数の意味があることがわかる。従って、格フレームは用言の用法別に構築する必要がある。

3. {母, 祖父, …} が {壁, 椅子, …} に {絵, コート, …} を かける

従来の研究では、コーパスから自動構築でカバレージの大きな格フレーム辞書を作る方法が提案されてきた。自動構築された格フレームの一つに京都大学格フレームがある (河原 and 黒橋, 2006)。この格フレームは Web 上にある約 16 億文の日本語テキストから自動構築されており、約 4 万用言に対応しているためカバレージが大きいと言える。しかしながら、格フレームを構築する過程で用言の用例をクラスタリングしているが、その結果が正確ではない。例えば、“挨拶”という用言の格フレームでは、言語を表す“日本語”及び“フランス語”が場所を表す“ステージ”が混ざった用例が生成されている問題がある。それとは別に、言語や場所が表す用例がそれぞれ生成されているため、不要な用例が生成されてしまっている。日本語の代表的な係り受け解析システムには KNP(河原 and 黒橋, 2007) や CaboCha(工藤 and 松本, 2002) があり、両方のシステムで京都大学格フレームが部分的に用いられている。これらのシステムで 3 のような文を係り受け解析すると、“カメラで”が“走っている”に係ると誤った結果になる。これは先ほど述べたノイズが影響していると考える。

そこで、本研究ではやさしい日本語辞書 (山本, 2017) の基礎語彙に限定して人手で構築した格フレーム辞書を提案する。やさしい日本語辞書は単純な高頻度語ではなく、あらゆる表現を可能にする語が登録されている。そのため、従来の研究のように、様々な表現に対応することができる格フレームを構築することができると考えられる。また、自動構築された格フレーム辞書は管理が不十分で多くのノイズが含まれ、自然言語処理での応用が難しい場合もある。しかしながら、やさしい日本語辞書の限られた範囲であるが、人手であるからノイズの少ない格フレーム辞書を構築することができる。また、高品質な格フレーム辞書は構文解析や機械翻訳などの様々な自然言語処理に応用しやすくなると考えられる。

2　対象の表現

本研究では、次のような語や表現に対して格フレームを構築する。対象としている用言は 591 語であるが、1 章で述べたように用言の多義性を考慮するため、格フレームの数は用言の数よりも多くなる。また、日本語には下記以外の格も存在するが、今回は最も一般的な格である“ガ・ヲ・ニ・デ格”を対象としている。格要素は名詞やサ変名詞などを合わせた 1155 語を主な対象としている。ただし、原則として“こと”や“もの”といった普遍的な語と“それ”や“これ”といった指示代名詞は、格要素の対象外としている。

- 動詞 (367 語) とサ変名詞 (221 語) を合わせた 588 語の用言に対して格フレームを構築する。
- 対象とする格は“ガ・ヲ・ニ・デ格”とする。
- 名詞 (912 語) とサ変名詞 (221 語)、代名詞 (22 語) を合わせた 1155 語を格要素の主な対象とする。

3　構築方法

格フレーム構築は作業者一人で以下のように行なった。

3.1　Web 日本語 N グラムの解析

Web 日本語 N グラムの解析では、各用言のどの格にどのような格要素が来るのかを調査し、頻度を求める。まず、3 gram の語が全て基礎語彙で構成されているものだけに着目し、“格要素, 格, 用言の表層形”となっている 3 gram を抽出する。例えば、“車, に, 乗る”といった 3 gram である。次に、抽出された 3 gram の頻度を計算し、用言と格要素の関係を格フレームとしてまとめる。

3.2　格要素のグルーピング

格要素のグルーピングでは、用言に来る格要素をカテゴリに置き換えるために行なう。まず、格要素の対象である 1155 語を同じ用言の同じ格に来る語が同じカテゴリに属するように人手でグルーピングする。グルーピングによって一つの格要素が複数のカテゴリに属する場合もあるが、一つの格要素しか属さないカテゴリは意味がないため作らない。次に、カテゴリが用言の格に来る頻度を求める。カテゴリの頻度はそのカテゴリに属する格要素の頻度の和をとったものを採用している。このように用言がとる格要素をカテゴリ単位で考えることで、Web 日本語 N グラムを解析しても分からなかった格要素にも対応することができると考えられる。

3.3　格フレームの構築・調整

格要素にカテゴリを取る格フレームを各用言の語義別に構築する。まず、3.1 節で構築した格フレームの格要素をその格要素が属するカテゴリに置き換える。しかしながら、このままでは自動構築と変わらず、不要なカテゴリ名が多く含まれている格フレームである。そのため、人手で不要なカテゴリや格要素を削除し、人間の知識により近い格フレームを構築することができる。さらに、やさしい日本語対訳コーパスの一部を係り受け解析し、格フレームに存在しない格要素があった場合は追加することで格フレームを改善する。これにより、作業者が想起できなかった用例にも対応できるようになる。

4　構築した格フレーム

本研究で作成したカテゴリは 84 個であり、格フレームの数は 621 個である。格フレームの数が用言の数よりも多いのは、用言の語義で格フレームを分割したためである。ここで、本研究で作成したカテゴリとそのカテゴリに属する語の一部を表 1 に示し、構築した格フレームの例を表 2 に示す。

5　係り受け解析による評価

本研究で提案する格フレームを用いて係り受け解析し、既存の係り受け解析システムである KNP と CaboCha との解析結果を比較する。やさしい日本語対訳コーパスのうち、格フレームの調整に用いなかった平易文 102 文を抽出し解析対象とする。それぞれの解析結果は作業者 1 人が人手で正しい係り受け解析が行われているかを判定する。KNP

表 1: 人手でグルーピングした結果であるカテゴリ及びそのカテゴリに属する格要素の例

カテゴリ名	カテゴリに属する語
ヒト (person)	私 (I), 父 (father), 友達 (friend), …
場所 (place)	海 (sea), 山 (mountain), 森 (forest), …
飲み物 (drink)	水 (water), ジュース (juice), 茶 (tea), …

表 2: 述語“あげる”の格フレーム例

述語	格	格要素のカテゴリ例 [頻度]
あげる (give)	が	ヒト [5,444], 敬称 [1,231], ヒト（役割）[149], …
	を	変化 [146,898], 文字 [49,015], 飲み物 [43,172], もの [27,879], …
	に	ヒト [97,091], 敬称 [26,082], 位置 [6,204], …
	で	方向 [2,925], ヒト [2,924], ただ [1,901], …
あげる (raise)	が	ヒト [5,444], 敬称 [1,231], ヒト（役割）[149], …
	を	ヒト（要素）[82,651], 飲み物 [43,172], もの [27,879], …
	に	方向 [12,694], 位置 [6,204], 家具 [2,664], …
	で	ヒト [2,924], ヒト（要素）[1,137], …

や CaboCha の解析結果には、本研究の対象外である格も解析されるが、それらの結果は今回の評価対象外としている。そのため、“ガ, ヲ, ニ, デ”格が正しく解析されているかに着目して評価を行なっている。

6 結果と考察

係り受け解析した結果より、評価文中 (102 文) に対象の格は 126 箇所であった。それぞれのシステムの評価結果は表 3 に示す。この結果より、制限された語彙の範囲ではあるが、構築した格フレームを用いた係り受け解析システムは既存の係り受け解析システムと同じくらいの性能があることがわかる。特に、本手法では格フレームの情報のみで係り受け解析をしており、係り受けの交差を禁止するといった制限は設けていない。だが、同程度の結果を得られている理由は、人手で構築した格フレームが高品質で有効に働いているためだと考えられる。

表 3: それぞれのシステムで解析した結果である解析対象箇所の正解数と適合率

手法	正解数（全 126 箇所）	適合率
Original	124	0.984
KNP	124	0.984
CaboCha	126	1.00

一方で、私たちの格フレームを用いた解析では、既存の係り受け解析システムで正しく解析できない文でも正確に解析することができることを確認した。例えば、次のような文がある。

1. 私たちはニュースで死んだ女性を知った。
2. 車がそんなに混んでいなければ問題ないでしょう。
3. 彼はいつも父のいないところで悪いことを言います。

例文 1 は、“ニュースで”が“死んだ”に係ると誤って解析されているが、正しくは“知った”に係る。これは、1 章の例文 1 と同様の間違え方である。また、例文 2 は“車が”が“ないでしょう”に係ると誤って解析されるが、正しくは“混んでいなければ”に係るべきである。同様に、例文 3 は“ところで”が“悪い”に係ると誤って解析されるが、正しくは“言います”に係る。これらの原因として考えられるのは、各動詞がとる格要素を適切に把握した格フレームが構築されていないことである。自動構築された格フレームは不要な格要素がノイズとして現れ、それにより誤った解析結果をもたらす。それに加え、解析する上での制約 (例えば、格要素と動詞の距離が近いほど係りやすい) が強すぎるため、語の並び替えが起こっている複雑な文章を適切に解析できないこともある。その点、私たちの格フレームを用いた解析では、厳しい制約を設けることなく正しい係り受け解析を行

える。従って、私たちの格フレームはノイズが少なく、実用的な解析においても有意義に働くと考える。さらに、格フレームは人手で構築しているため、修正が容易で品質を向上させていくことができる。また、格要素のカテゴリにやさしい日本語辞書以外の単語を追加することで、カバレージを大幅に向上させることができる。例えば、“ヒト”というカテゴリに人名や人を表す単語を自動で追加することが挙げられる。このようにカバレージを向上させるための作業は自動できるため、人手で作業が必要な部分と自動で行なう作業を的確に見極めていくことが重要である。

7 関連研究

日本語の格フレームを構築する研究は、河原らの研究がある。この研究では Web 上の約 16 億文の日本語テキストから自動的に構築しており、約 4 万の用言が登録されている。その格フレームは構文解析や意味解析で用いられている (河原 and 黒橋, 2007)。また、多言語の研究では semantic role labeling (SRL) がある。SRL は文の内容を抽出することであり、格フレーム辞書を構築することであると考えることもできる。(Jin et al., 2017) は中国語の SRL を高品質な格フレームで改善する手法を提案している。改善に有効な格フレームが表層格と深層格のどちらであるかについても言及されている。この研究から、高品質な格フレームを用いることで、格フレームを用いた解析の精度が向上することがわかる。さらに、深層格フレームの方が表層格フレームよりも豊富な知識を持ち、解析に優位に働くこともわかる。さらに、自動的に行なっているため必ずノイズが発生し品質が低くなるが、比較的高品質な格フレームを用いると性能が向上することがわかっている。したがって、人手で構築した高品質な格フレームは構文解析や SRL の改善に大いに寄与することが期待される。

8 おわりに

本研究ではやさしい日本語辞書の範囲で格フレームを人手で構築した。その結果、588 語の用言に対して、621 個の格フレームを構築することができた。私たちの格フレームを用いて係り受け解析をした結果、既存の係り受け解析システムと同程度の解析精度を得ることができた。さらに、既存の解析システムが間違う複雑な文を正しく解析することができた。今後の研究では、他の格を格フレームに追加し解析できる範囲を広げ、格解析にも応用したい。また、格要素となりうるやさしい日本語辞書以外の単語をカテゴリに追加することで、カバレージを大きくしていきたい。

謝辞

本研究は、平成 27〜31 年科学研究費補助基盤 (B) 課題番号 15H03216、課題名「日本語教育用テキスト解析ツールの開発と学習者向け誤用チェッカーへの展開」、及び平成 29〜31 年科学研究費助成事業挑戦的萌芽課題番号 17K18481、課題名「やさしい日本語化実証実験による言語資源構築と自動平易化システムの試作」の助成を受けています。

参考文献

Jin, G., Kawahara, D., and Kurohashi, S. (2017). Improving chinese semantic role labeling using high-quality surface and deep case frames. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 568–577. Association for Computational Linguistics.

河原, 大. and 黒橋, 禎. (2006). 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. 情報処理学会研究報告自然言語処理 (NL) , 2006(1):67–73, jan.

河原, 大. and 黒橋, 禎. (2007). 自動構築した大規模格フレームに基づく構文・格解析の統合的確率モデル. 自然言語処理, 14(4):67–81.

工藤, 拓. and 松本, 裕. (2002). チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, 43(6):1834–1842, jun.

山本, 他. (2017). やさしい日本語対訳コーパスの構築. 言語処理学会第 23 回年次大会, pages 763–766.