箱庭言語処理のための格フレーム辞書構築の意義

Significance of constructing case frame dictionary in 2k-word language space

長岡技術科学大学准教授
山本　和英

豊橋技術科学大学大学院工学研究科博士後期課程システム情報工学専攻修了。博士(工学)。1996年～2005年(株)国際電気通信基礎技術研究所(ATR)、2002年～現在まで長岡技術科学大学、現在准教授。自然言語処理の研究に従事。言語処理学会理事、アジア太平洋機械翻訳協会理事。

yamamoto@jnlp.org

1　はじめに

本稿では日本語の構文解析について論じる。現在の自然言語処理の潮流は、人手による事前知識の構築を否定し、大量のコーパスから何らかの知識を自動獲得しようという流れが続いている。タスクによっては（人手で構築した辞書による）形態素解析すら不要とする立場もある。私はこのような最近の流行に対して大いに危機感を持っており、これについては［山本 2017b］で述べた。この流れは今後も主流であり続けるであろうが、これとは一線を画して、我々のできる範囲で言語知識の構築を続けていかなければならない、と私は考えている。

日本語処理において構文解析（格解析、係り受け解析）は重要な要素技術の一つと言われ続けてきた。しかし、産業応用という観点から、実際に日本語において係り受け解析が使われている応用システムは、私の見る限りそれほど多くない。もう少し強く、ほとんどないと言っていいかもしれない。また、このような状況と関係しているかどうかは不明だが、構文解析の研究も決して多いとは言えないのが現状認識である。

日本語の係り受け解析については、KNP と CaboCha という二つの解析システムが存在し、どちらも京都大学格フレームが情報源として使われている。この格フレームは日本語において利用されている唯一の格フレーム辞書と言ってもいい[1]。日本語において類似する言語資源が存在しないため、京都大学格フレームの学術的な貢献は非常に大きいが、その一方で改善の余地は大いにある。最大の問題は、品質の問題である。

京都大学格フレームは大規模な格フレームであり、詳細は割愛するが半自動的な手法によってできるだけ品質を高める工夫を行っている［河原 2007］。構築対象が約４万用言と大規模であるため全構築データを人手で確認するのは現実的に不可能に近く、また万が一行うにしても膨大なコストが必要である。その結果として半自動による構築を選択するのは極めて合理的であり、また現実的であると思う。しかし、その結果として依然ノイズが多数含まれた言語資源が作成され、これを情報源として解析モデルを構築すると、解析精度はある時点で頭打ちになり、それ以上の解析精度は基本的に得られない。現在はまさにこの状態なのではないかという認識を持っている。そうかと言って、現在の京都大学格フレームに人手で手を加えるというのは、構築時と全く同じ理由で合理的ではなく、よほどのことがない限りこのような膨大で地味な作業はやりたくない。

この問題に対する方策の一つとして、私は以下のように考えた。格フレーム全体を人手で構築するのは不可能に近くても、その一部であれば構築可能なのではないだろうか。すなわち、大規模でやや品質が劣る格フレームを半自動で構築するよりも、小規模で高品質な格フレームを手作業で構築して、これと現実の大規模語彙の間の写像を自動的に行う何らかの仕組みと組み合わせるほうが、解析精度の高い構文解析のために有益なのではないか、という提案である。この考え方は、格フレームに限らず、大規模な言語資源を必要とする多くの自然言語処

1　もう一つ、日本語語彙大系の構文体系が存在するが、研究としてほとんど使われていないのは非常に残念に思う。

理タスクにおいて現実的で効果のあるアプローチではないかと私は考えている。

そして、この小規模言語空間として、我々が「やさしい日本語」と称する言語空間を用いる。「やさしい日本語」とは、日本語初学者のための情報伝達手段として、英語等へ翻訳するのではなく平易な日本語を用いることを意味するが、「平易さ」について語彙的、文法的に明確な定義は存在しない。そこで我々は独自に２千語を選定し、この（固有名詞等を除いて）２千語のみで記述された言語表現を「やさしい日本語」として定義した[山本 2017a]。そして、選定した２千語はある種の中核的表現と捉えることもできるので、我々はこの２千語で記述された文に対する解析を「箱庭言語処理」と呼んでいる。なお、箱庭言語処理の意義、及びやさしい日本語との関係は昨年度の Japio YEAR BOOK［山本 2017c］において議論した。

本稿では、以上のような問題意識から、やさしい日本語２千語に対する格フレームの構築について述べる。なお、詳細については［角張 2018］を参照されたい。

2 やさしい日本語格フレーム辞書の構築

2.1 構築上の問題

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

本研究ではやさしい日本語辞書[山本 2017a] 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。

2.2 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、前節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 1,155 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

格フレーム構築は本研究室学部生（作業当時）の角張竜晴さんが一人で以下のように行なった。

3.1 コーパスから格関係事例の収集

テキストコーパスから名詞（例:車)、格助詞（例:に)、用言（例：乗る）のすべてがやさしい日本語 2 千語で構成されている三つ組を収集する。この際、構文解析器は使わずに単語が連続して出現するもののみを擬似的に格関係があると見なして収集する。収集には Web 日本語 N グラムを使用する。

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 1,155 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

本研究で作成した名詞カテゴリは 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。

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], …

構築した格フレーム、及び名詞カテゴリを観察したところ、いくつか興味深い知見が得られた。例えば、表 1 で例を示した名詞カテゴリは、いわゆるシソーラスなどの分類と同様（すなわちこのような作業は不要）ではなく、格フレーム構築に必要な「ちょうどいい粒度」が必要である。例えば、人間に関する分類はすべて「ヒト」カテゴリでいいと思いがちであるが、作業の結果下記の

5 分類が必要である。これらは本件のような格フレーム構築作業を行って初めて分かる分類粒度であり、貴重な言語資源と考える。

なお、格フレーム構築の副産物として得られたこの名詞分類は、今回のような格フレーム構築の際の汎化のみならず、様々な名詞汎化作業において最適な汎化粒度ではないかとも考えている。この点についてもいずれ検証を行いたい。

（人に関する分類項目と所属する単語の例）
- ヒト：父、女、老人、友達、相手
- ヒト（役割）：ドライバー，ファン，プロ，モデル，上司
- ヒト（組織）：集団、家庭、企業、民族、政党
- ヒト（数）：みんな、一人、二人、全員、自分
- ヒト（要素）：手、舌、神経、肌、臓器

5 おわりに

本稿では我々が進めている「箱庭言語処理」の一環として、完全に人手で構築したやさしい日本語格フレームについて述べた。小規模ではあるが、我々が選定した「やさしい日本語」語彙である 2 千語に対して、人手で動詞と名詞の格関係を記述し、辞書としてまとめた。また、この辞書を一般に公開した。

SNOW D16：やさしい日本語格フレーム
http://www.jnlp.org/SNOW/D16

この辞書は、人手で構築したものではあるがまだ完全なものとは認識していない。このため今後は時間をかけて、誤りを見つけ次第修正および拡充をしていきたいと考えている。

このような小規模高品質の格フレームが本当に現実の日本語文に対する係り受け解析、格解析に有効なのかについては、我々で小規模な解析実験を行った結果では、（やさしい日本語語彙に存在する語彙に対して）すでに同等程度であるとの結果が得られている[角張 2018]。しかし、この実験はまだ小規模かつ限定的であり、特にやさしい日本語として選定した 2 千語以外の語彙に対して適用するためにどのような汎化を行い、その結果どの程度の解析精度が得られるかについては何も検証できていない。これについては今後検証を進めていきたい。

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

[角張 2018] 角張竜晴，山本和英．やさしい日本語格フレームの構築による係り受け解析．言語処理学会第 24 回年次大会，pp. 777-780（2018）

[河原 2007] 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・格解析の統合的確率モデル．自然言語処理，Vol. 14, No. 4, pp. 67-81（2007）

[山本 2017a] 山本和英, 丸山拓海, 角張竜晴, 稲岡夢人, 小川耀一朗，勝田哲弘，髙橋寛治．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，pp. 763-766（2017）

[山本 2017b] 山本和英．知識を書こう．自然言語処理，Vol. 24, No. 4, pp. 521-522，言語処理学会（2017）

[山本 2017c] 山本 和英．箱庭言語処理―2 千語の言語空間における言語処理の意義―．Japio YEAR BOOK 2017, pp. 342-345，日本特許情報機構（2017）