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―

natural language processing 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、及び日本語教育（ツール作成）の研究に従事。

yamamoto@jnlp.org

1　はじめに

本稿では「箱庭言語処理」という概念を提唱する。これは本研究室で現在進めている一連のプロジェクトの総称である。一般の言語空間の部分集合である言語空間（我々は「やさしい日本語」と呼んでいる）を定義して、その空間内において様々な自然言語処理のタスクに取り組むという試みである。後述するように、このような試みを行うことによって従来一般の言語空間において十分な精度が得られずにいた言語処理タスクについて解決の糸口がつかめるのではないかと我々は見込んでおり、またやさしい日本語という特殊な言語空間において様々な処理が実現できれば単独でも実用性が見込めるという点でいわゆるトイモデルとは異なる。

本稿では２章でこのプロジェクトを進める上で密接な関係にあるやさしい日本語及び自動平易化について説明を行ったあと、３章でやさしい日本語と箱庭言語処理の関係について述べる。４章では箱庭言語処理にどう取り組むかについて我々の目論見などについて総括的な議論を行い、５章で本稿のまとめを行う。

2　やさしい日本語

「やさしい日本語」は近年重要性を増している考え方である。日本語初学者、子供、障害者などに向けた情報提供の一手段として重要であり、例えば関連する書籍もいくつか出版されている（例えば [庵 2013]）。これに伴って（日本語の）自然言語処理においても自動平易化研究が行われている（例えば [梶原 2015]）。

本研究室では一般の日本語文を入力し、やさしい日本語へ自動で書き換える自動平易化のための様々な研究を進めている。まず手始めとして、日本語文とやさしい日本語文が対応している「やさしい日本語コーパス」を構築した（[山本 2017]）[1]。このコーパスのすべての文は（固有名詞を除いて）２千語の語彙で書き換えられており、コーパスの規模は５万文である。

従来、平易な日本語を用いたコーパス作成に関する関連研究として [李 2013] がある。ここでは約４万文の公的文書（市役所から市民へのお知らせ文書）に対して文単位でやさしい日本語に書き換えた。このコーパスの大きな特徴はほぼ実在の文書に対して日本語教師が書き換えを行っている点で、長文や難解な概念を含んだ文に対しても分かりやすく書き換えられていたり、場合によっては冗長な言い回しという理由で全文削除されていたりする。また、様々な分野、話題を含んでいる。この結果、語彙数や文長、分野などの理由からこのコーパスを用いた処理は難解と判断し、より処理しやすい（つまり現象として簡単化した）コーパスを作ることにした。以上をまとめて本研究と比較した表を表１に示す。

我々は、書き換えを行う対象となる原テキストとして、small_parallel_enja: 50k En/Ja Parallel Corpus for Testing SMT Methods[2] を採用した。このテキス

1　後述するようにこのコーパスは元々日英対照なので、我々の作業によって３対（日本語、英語、やさしい日本語）が文単位で対応している。

2　https://github.com/odashi/small_parallel_enja

表１　日本語の二つの平易化コーパスの比較

	[坐 2013]	本研究
文数	42,274 文	５万文
使用語彙	(6,000 語相当)	２千語
作業者	日本語教師	一般（学生）
分野	公的文書	(非限定)
平易化	３種類（逐語訳/意訳/要約）	１種類
英訳	なし	あり

トは田中コーパス[3]の部分集合で日英機械翻訳のために作成された小規模対訳コーパスである。このコーパス中にある５万文すべてをやさしい日本語に変換することとした。

このコーパスの構築にあたって、やさしい日本語も同時に定義した[4]。まず２千語という語彙規模を決め、コーパスに対する書き換え作業を進めながら２千語の語彙を同時に洗練していく、という手順で作業を行った。また、我々自身の作業の効率化のために、「やさしい日本語チェッカー」[5]というツールも作成した。このツールの概観を図１に示す。このツールは、入力した文に含まれる単語がそれぞれ（暫定的に定義している）２千語に含まれているかどうかを表示するツールである。

図１　やさしい日本語チェッカー

さらに、現在はこのコーパスを拡張する作業も進めている。第一期で作成した５万文は研究室内で学生が分担して書き換え作業を行ったが、第二期では上記田中コーパスで書き換え作業を行っていない文を対象にしてクラウドソーシングを用いて書き換え作業を進めている。この結果得られたデータの特性やクラウドソーシングによる作業品質などについては書き換え作業終了後に調査を進めていく。

また、第一期で得られたコーパス５万文を用いて自動平易化の研究を現在進めている。この研究成果については、今後学術的な場で報告したいと考えている。

3　やさしい日本語と箱庭言語処理

次に、やさしい日本語と箱庭言語処理との関連について議論する。

やさしい日本語の研究は自動平易化の実現が直接の目的である。しかし、今回我々が定義した「やさしい日本語」という言語空間は、一般の日本語と比較すると使用される語彙のみが異なり、その他の言語現象には一切の制限を加えていない。これは自然言語処理の観点で考えると、解かなければいけない問題は（未知語収集など語彙数に極度に依存する一部問題を除いて）ほぼ残っており、語彙関連の問題に関しても（使用した２千語の範囲での）語義曖昧性解消問題などの問題は依然としてそのままである。すなわち、我々が定義したやさしい日本語という言語空間は、使用語彙のみが制限されているが自然言語処理の問題として何も簡単化されていない特殊な言語空間だと捉えることができる。

我々は、自動平易化とは別の観点からこの特殊な言語空間に着目した。すなわち、この言語空間に対して言語処理問題を解くことによって、自然言語処理の技術の進展が可能なのではないかと考えた。一般の言語空間は語彙の規模が大きすぎて、どのような手法で解こうにも多大なコストがかかる。何らかの言語資源を作ろうにも膨大な時間がかかるし、そもそもそのようなコストを費やす価値があるのかどうかの判断ができないので大規模化のリスクを取ることが難しい。コーパスを用いて自動処理しようとしても膨大な計算量と膨大な記憶容量がかかり、さらにそれが組み合わせで効いてくると（恣意的な語彙削減をしない限り）現実に処理するのは難しくなる。仮に可能だとしてもモデルの改良に時間がかかり、十分な精度が得られない場合にそれがモデルの問題なのかデータの過疎性の問題なのか、明確に分離できない。

3　http://www.edrdg.org/wiki/index.php/Tanaka_Corpus

4　本研究で言う「やさしい日本語」は使用語彙を制限しただけなので、本当の意味での「やさしい日本語」ではないが、便宜上この呼称を用いている。

5　http://www.jnlp.org/SNOW/S13

以上より、どのようなアプローチで自然言語処理問題を解くかに関わらず、やさしい日本語の言語空間を対象にして研究を進めることで従来の大語彙を対象にした処理では隠れて見えなかった問題点や解決策が明らかになる可能性がある。我々はこのように考えて、自動平易化以外についてもやさしい日本語を対象にして研究を進めることにした。また、この研究の総称を箱庭言語処理と呼ぶことにした。

箱庭は、単に現実世界を切り取った模型というだけでなく、詳細にわたってとても精巧に作られているという意味合いがある。このことから、現実を極端に簡単化した世界ではないという思いでプロジェクト名に箱庭という用語を使うことにした。

4 箱庭言語処理に対する我々の取り組み

箱庭言語処理とは、使用する語彙を制限した状況において行う言語処理のことである。前述したように、自然言語処理の部分問題と考えることもできる。この問題に対して、我々はどのように取り組むのかについて述べる。

自然言語処理の高度化のためには、何らかの形でより高度な語彙資源を整備する必要があると我々は考えている。現在のように、言語資源として形態素解析辞書とシソーラスのみを使用している状況では明らかに言語情報として不足している。現在主流となっているコーパスを用いた自然言語処理においては様々な知識獲得の試みが続けられている。ここでは、基本的には単語の出現文脈が情報源で、必要に応じてアノテーションを行った上で何らかの言語知識を獲得する。しかし、テキストの情報だけで人間の持つ言語情報や知識を獲得することはおそらく不可能で、仮に可能だとしても非常に効率が悪い。このため、コーパスを用いた言語処理は有益だとしても、これのみですべての処理を行うのは賢明ではないと考えるのが我々の立場である。

このように、何らかの知識（規則、パタン、辞書など。本稿では、予め作成する言語処理に必要な静的情報を知識と呼ぶ）を構築する必要性があることは感じているが、その一方でこれら言語知識は簡単に構築できるものではない。一般に、言語知識は大規模となるので、有効性が確認できない状況において作り始めるには大きなリスクを伴う。

そこで、本研究では「箱庭言語処理」という考え方を提唱している。これは、語彙を限定した言語空間を作り、この中で言語資源を作成してその有効性を確認しようという考え方である。これの意義は以下の２点にある。

（A）「箱庭」モデルを単独で使用する

ここでの箱庭言語処理として、２節で説明してやさしい日本語２千語の言語空間をそのまま使用する。２節で述べたように、我々はすでに２千語で書かれた５万文のコーパスを所有しており、現在さらにこれを拡張すべくクラウドソーシングでテキストの書き換え作業を進めている。仮に田中コーパス全文である 16 万文まで収集できれば、一定規模のモデルの検証や機械学習が可能になるのではないかと考えている。

２千語で書かれた文は語彙制限があるため原文と同じ意味を表現できないことも多いが、その一方である程度は原文に近い意味を表現できていることも確かである。よって、仮にこの２千語で書かれた日本語に対して様々な自然言語処理の問題が解けるようになれば、少なくともこの言語空間においてはいろいろな自動処理が実現することを意味する。現在の自然言語処理の性能が依然として限定的であることを考慮すると、限られた状況であるにせよ２千語を使って自然言語を記述すれば機械で誤りのない処理ができるという状況を創り出せるかもしれない。

（B）箱庭モデルと現実世界との写像

次に、箱庭モデルはある種の部分空間であるが、もしこの空間において問題が解けたのであれば、現実の自然言語を解く方法はいくつかある。

一つは、空間の拡張である。現在は我々が選定した２千語のみを用いた言語空間を考えているが、将来はこれを例えば６千語、さらには２万語のように拡張していくことで徐々に現実の自然言語に近づいてくる。２千語のみで構築した何らかの語彙知識も、２千語でうまく機能しているという結果が得られているのであればこれを大規模化していくことにそれほどリスクはないであろう。

もう一つは、やさしい日本語空間と現実との写像を行うことである。箱庭モデル自身を拡張するという考え方を採用するのでなく、このモデルはこのままある種の意

味空間と考え、現実の言語空間との間で何らかの写像を行うことができれば、箱庭モデルをそのまま内部処理として用いることができるかもしれない。

5 おわりに

本稿で考えているような、ある種制限された状況における自然言語処理のアプローチは私の知る限り聞いたことがない。しかし、自然言語処理は様々な問題が複雑に関係していることが明らかになっている以上、このような考え方で研究を進めていくことも必要なのではないだろうかと我々は考えている。

箱庭言語処理の利点は、決して人工的な言語空間でもなくトイモデルでもないという点、及びその処理に一定の実用性がある点の2点にある。今後はやさしい日本語への平易化処理の研究を進めると同時に、やさしい日本語空間における構文解析や意味処理など、様々な言語処理の高度化に挑戦していきたい。

謝辞

本研究は、平成 27～31 年科学研究費補助金 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

[庵 2013] 庵 功雄, イ ヨンスク, 森 篤嗣（編集）.「やさしい日本語」は何を目指すか：多文化共生社会を実現するために. ココ出版（2013）

[梶原 2015] 梶原 智之, 山本 和英. 語釈文を用いた小学生のための語彙平易化. 情報処理学会論文誌, Vol.56, No.3, pp.983-992, 情報処理学会（2015）

[杢 2013] 杢 真奈見, 山本 和英.「やさしい日本語」変換システムの試作. 言語処理学会第 19 回年次大会, pp.678-681（2013）

[山本 2017] 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）