自然言語処理の基礎

工学博士　奥村　　学 著

コ ロ ナ 社

ま　え　が　き

筆者が自然言語処理という言葉を初めて知ったのは，大学の学部 3 年生のころであった。筆者が高校生のころはまだパソコンなどというものはなく，コンピュータというものはまだ世の中で一般的ではない時代に，私はなにを思ったか「コンピュータって面白そう」と勝手に思い込み，「情報工学科」に入学した。入学したものの，大型計算機でのプログラミングには必ずしもなじめず，四苦八苦し，卒業研究でもなにをテーマにしようか迷っていた。そのとき本で人工知能，とりわけわれわれの言語を理解するコンピュータの話を読んでこれだと思ったわけである。

筆者にとっての幸運は，4 年生になった年に，当時出版された本で，自然言語処理の草分けとして名前のあった田中穂積先生が，東工大に着任されたことであった。いうまでもなく，卒業研究では田中先生の研究室を希望し，以後，現在までの 26 年間，自然言語処理というテーマで研究開発を行うことになった。

筆者の卒業研究は，（構文解析で用いる）日本語の文法を作成するというテーマであった。1980 年代の自然言語処理はまだ小規模なシステムが多かったが，1 章で説明するように，その後，筆者自身も自然言語処理技術の歴史的変化，大きな技術的な変革を直接経験することになる。筆者自身の研究の立場も，基礎的な自然言語を理解するという研究から，WWW 上のテキストデータを対象とした応用研究の割合が少しずつ大きくなっていった。26 年間の自然言語処理との付合いの中では，研究開発した技術を基に会社を立ち上げるという貴重な経験もすることができた。いまや自然言語処理技術を基に会社だって設立できるのである!

また同時に，長年，教員として自然言語処理を若い学生さんに教えてきて，今回自然言語処理の教科書を書くという機会を得た。この教科書は，学部の 3,

4 年の学生さんが初めて自然言語処理について学ぶことを念頭に置いて書いたものである。また，半期の講義で扱える程度の内容を盛り込むように配慮したつもりでもある。伝統的に重要であると考えられている，人間の作成した知識を用いた自然言語処理手法とともに，コーパスから獲得した確率などを用いた，近年の自然言語処理手法についてもふれるように努力した。また，実用に近い自然言語処理技術の応用についても可能なかぎり説明するように努めた。

本書だけで自然言語処理のすべてがわかるということは，残念ながらないと思うが，本書の中で紹介している他の書籍なども参照して，さらに自然言語処理の理解を深めて下されば幸いである。

非常に残念だが，昨年 7 月われらが恩師である田中先生は，若くしてお亡くなりになった。本書を端緒として，自然言語処理って面白い! 自然言語処理って役に立つんだ! といったように，自然言語処理の面白さ，自然言語処理の可能性に若い読者が気づき，自然言語処理の世界にふれてくれるようになるなら，田中先生から自然言語処理の面白さ，難しさ，そして実用化の可能性について教えていただいた筆者にとっては，望外の喜びである。われわれの後をさらに引き継ぎ，自然言語処理の分野を引っ張っていってくれる若い研究者が，本書の読者から出てくれることを心から望んでいる。

この本を出版するにあたり，まずコロナ社に感謝します。なにかと遅れがちな原稿作成の過程で，つねに叱咤激励をいただきました。出版の日を迎えられたのは皆様の激励があったからこそともいえます。また，田中穂積先生がいらっしゃらなければ今日の自分はなく，また本書はなかったといえます。ご冥福をお祈りするとともに，学部の 4 年生からずっとご指導，ご助言，ご支援下さったことに心から感謝いたします。ありがとうございました。また，本書の草稿に貴重かつ適切なコメントを下さいました，東京工業大学情報理工学研究科の徳永健伸教授，北陸先端科学技術大学院大学情報科学研究科の白井清昭准教授，広島市立大学情報科学研究科の難波英嗣准教授にも心から感謝いたします。

2010 年 8 月

奥 村　学

目　　次

1. 自然言語処理概論

2. 辞書とコーパス

3. 形態素解析

4. 構 文 解 析

5. 意 味 解 析

6. 文 脈 解 析

7.　自然言語処理の応用

1 自然言語処理概論

本章では，われわれ人間が日常的に話し聞き，読み書きしている言語をコンピュータ上で処理する技術である自然言語処理の概要について説明する。特に，自然言語処理の歴史，応用について説明する。

1.1 自然言語処理とは

自然言語（natural language）とは，日本語，英語，フランス語など，われわれ人間が日常的に話し聞き，読み書きしている言語のことをいう。歴史的には，コンピュータにおける言語といえば，C, Ruby, Perl などのプログラミング言語，すなわち人工言語のほうが一般的であったため，人工言語と区別するため，ただ単に「言語」というのではなく「自然言語」と呼ぶこととなった。この自然言語をコンピュータ上で扱う技術を**自然言語処理**（natural language processing）という。

自然言語処理は，大きくつぎの二つの立場に分けることができる。

- コンピュータに「ことば」を理解させる。
- コンピュータ上で「ことば」を処理する。

次節以後で説明するように，自然言語処理の研究が始まった当初から，自然言語処理は，前者の立場に立って研究されてきたといってよい。いわゆる伝統的な人工知能（artificial intelligence）的研究であり，コンピュータ上に知的な能力をもつものを実現したいという動機づけで研究されてきた。

一方で，1990 年代に入ると，インターネットが急速に普及するのに伴い，

WWW 上などに膨大な量のデータが蓄積されるようになり，その中には，テキストの形で蓄積されるデータもかなりの量を占めるようになった。このような形で，コンピュータ上には大量のテキストデータが蓄積されるようになり，このテキストデータをデータとして加工することで，なんらかの有用な情報を取り出したり，大量のデータを有効に活用する技術が開発されるようになってきた。これが後者の立場の自然言語処理といえるだろう。後者の立場の自然言語処理については，7 章で詳しく説明する。

現在では，この二つの立場の研究が相補的に働き，自然言語処理の研究開発が活発に行われ，技術が深化するとともに，有用なシステムも続々と新たに生まれてきている。

1.2　自然言語処理の歴史

自然言語処理に関する研究開発は，コンピュータの開発とほぼ同時に開始されたといってよい。コンピュータが 1940 年代に出現したのと時(とき)を同じくして，機械翻訳システムの構想が生まれたというのは，今日から考えると驚くべき事実ではないだろうか。情報検索の歴史も同様に古く，1950 年代には情報検索の研究が始まっている。日本では，1970 年代に仮名漢字変換機能をもつワードプロセッサが登場し，以後，仮名漢字変換技術は，われわれにとって最もなじみのある自然言語処理の応用技術として用いられている。

1990 年代に入ると，インターネットが急速に普及するのに伴い，WWW 上には膨大な量のテキストデータが蓄積されるようになり，これらの膨大なテキストデータを有効に利用するための技術開発も進み，ユーザが欲しい情報を検索する情報検索技術や，逆にユーザが不要と考える情報をフィルタリングするフィルタリング技術，膨大なテキストデータの蓄積から，「おもしろい」情報を発掘しようとするテキストマイニング技術などが研究開発されるとともに，実用化されるに至っている。

一方で，自然言語処理の手法も同時期に大きな転機を迎えることになる。コ

ンピュータの HDD が安価になり，非常に大容量のテキストデータを保存できるようになり，また大量のテキストデータが容易に入手できるようになったことに伴い，それらの大量のテキストデータを「コーパス」として蓄積し，そのコーパスから自然言語処理に用いる知識として規則や確率を獲得し，それらを用いた自然言語処理の手法が生まれるようになってきている。本書では，2 章でまずコーパスについて説明し，その後の章では，個別の解析技術においてどのようにコーパスが用いられるかについて説明する。

1.3 自然言語処理はなににに使えるか

本節では，自然言語処理の応用について，典型的なものをいくつか紹介しよう。前節で説明したように，自然言語処理の研究の当初から研究開発が行われていた**機械翻訳**（machine translation）は，その代表例といってよいだろう。機械翻訳の手法については，7.1 節で簡単に説明する†。

また，これも前節で紹介したが，いまやわれわれにとってはなくてはならない入力手段である仮名漢字変換も，自然言語処理技術の典型的な応用の一つであろう。この技術のおかげで，キーボードから入力したローマ字あるいは仮名文字列は単語に分割されて漢字に変換され，われわれは現在何不自由なく日本語のテキストを入力できるようになっている。

もう一つ，わかりやすい自然言語処理の応用例を挙げよう。われわれが日常用いている言語でコンピュータやロボットとコミュニケーションができるようになると，よりそれらの機器はわれわれにとって親しみのあるものになる可能性がある。このようなコンピュータの機能は，対話システム，自然言語インタフェースなどと呼ばれる。これらのシステムの研究開発は，自然言語処理の歴

† 自然言語処理の書籍に，‘Time flies like an arrow.’（「光陰矢の如（ごと）し」）という文がよく例として登場するが，これは，古典的な機械翻訳システムが「時蝿（ときばえ）は矢を好む」と見事に誤訳する（「誂（ことわざ）」の翻訳というのは本来難しいのだが）ことにちなんでいると思われる。

史の中でかなり初期の段階から行われている。

少し古い話になるが，ここで，歴史上に名前の残る二つの対話システムについて説明しておこう。一つは，1966 年にワイゼンバウム（Joseph Weizenbaum）が開発した **ELIZA**（「イライザ」と発音する）である[56)†1]。ELIZA は，ユーザがコンピュータ上でチャットで対話できるカウンセラ役をすることができる対話システムである。**図 1.1** に ELIZA の対話例を示す。図では，全部大文字で表記されているのがシステムの発話である。英語なので難しいかもしれないが，図の対話例を見て，どのように感じられるだろうか。比較的自然に対話が進んでいるように感じられた読者もいるのではないだろうか。

ELIZA は当時自然な対話が行え，**チューリングテスト**（Turing test）[†2]にパスするということで注目を集めたが，さらに驚かされたのは，このシステムが，きちんとした自然言語処理を行っておらず（1960 年代であるから無理はないが），単にユーザの発話に含まれる特定の単語列と照合するパターンを用意しておき，照合したパターンに応じて応答を出力するだけのものであったことである。例えば

ユーザの発話	システムの応答
i am X	how long have you been X?
..	..

というパターン（パターン中 'X' は変数を表す）をもっていれば，'i am unhappy' というユーザの発話とパターンの左側が照合するので，'how long have you been unhappy?' というシステムの応答を出力できるというわけである。

ELIZA に類する対話システムは現在でも広く世の中に存在する。例えば，Emacs/Mule から M-x doctor で実行可能なプログラムがそうであるし，日本

†1　肩付数字は，巻末の引用・参考文献の番号を表す。

†2　チューリングテストとは，コンピュータの原型を考えたチューリング（Alan Turing）が 1950 年に提唱した，コンピュータの知能テストである。コンピュータに知性があるかどうかを判断するのに人間の被験者を用い，その被験者に端末を通して対話させる。人間の被験者が対話相手を人間と間違えるようならば，そのコンピュータは知性があると判断する。現在でも，チューリングテストを行うコンテストがレーブナー賞（Loebner prize）という名前で行われている（`http://www.loebner.net/Prizef/loebner-prize.html`）。なお，本書に掲載する URL は執筆当時のものであり，変更される場合がある。

Men are all alike.
IN WHAT WAY
They're always bugging us about something or other.
CAN YOU THINK OF A SPECIFIC EXAMPLE
Well, my boyfriend made me come here.
YOUR BOYFRIEND MADE YOU COME HERE
He says I'm depressed much of the time.
I AM SORRY TO HEAR YOU ARE DEPRESSED
It's true. I am unhappy.
DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY
I need some help, that much seems certain.
WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP
Perhaps I could learn to get along with my mother.
TELL ME MORE ABOUT YOUR FAMILY
My mother takes care of me.
WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU
My father.
YOUR FATHER
You are like my father in some ways.
WHAT RESEMBLANCE DO YOU SEE
You are not very aggressive but I think you don't want me to
notice that.
WHAT MAKES YOU THINK I AM NOT VERY AGGRESSIVE
You don't argue with me.
WHY DO YOU THINK I DON'T ARGUE WITH YOU
You are afraid of me.
DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
My father is afraid of everybody.
WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER
Bullies.
DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR
BOYFRIEND MADE YOU COME HERE

図 1.1 ELIZA の対話例（文献 6) より引用）

で「人工無脳」と呼ばれていたプログラムは，この ELIZA が起源であるということができる。最近では，Twitter などに存在するボット（bot）と呼ばれるプログラムも同様のものである。

もう一つの対話システムは，1972 年にウィノグラッド（Terry Winograd）が開発した **SHRDLU**（シュルドゥルと発音する）である[57)†]。SHRDLU は図

† 英語の文字を頻度順に並べると
ETAOINSHRDLU...
となるところなどから，このシステムの名前が付けられたとされている。

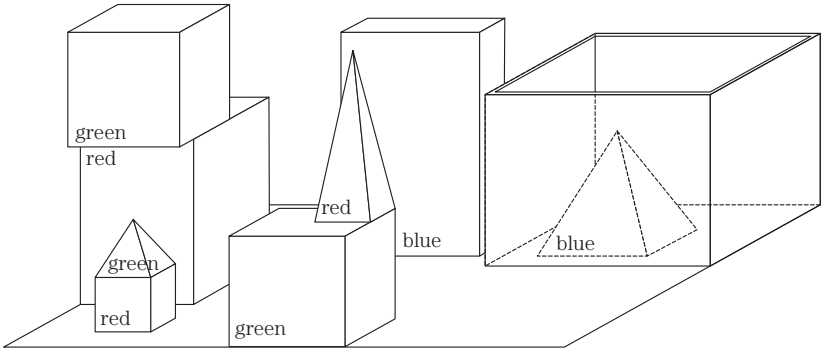

図 1.2　SHRDLU の積み木の世界

1.2 に示したような積み木の世界に存在するロボットアームに対して指示を出すことのできるシステムである。

仮想的につくられた積み木の世界にはロボット（アーム）があり，人間の自然言語による指示を理解し，積み木の世界を操作し，モニタ上で積み木の世界が変化する様子を表示する。例えば，図 1.2 の状況において

Pick up the blue box on the table.

と指示をすると，ロボットアームがテーブルの上の青の箱をもち上げ

Put it on the green box on the red box.

と指示すると，もっていた箱を（赤の箱の上の）緑の箱の上に置いてくれる。

指示を出した人間は，システムが言語による指示を理解したということを，モニタ上での世界の変化により（自分の出した指示のとおり世界が変化することで）確認することができる。SHRDLU は，積み木の世界という限定はあるものの，自然言語の指示を理解するコンピュータというものを明確に示した最初の例といえる。

最後に，自然言語処理に関連する，われわれにとってなじみ深いもう一つの応用として，Google，Yahoo! などに代表される検索エンジンを挙げておこう。検索エンジンの根幹をなす技術である情報検索は，自然言語処理とは別に研究開発が始まったが，いまやこの二つの技術は非常に近い位置にあるということ

索　　引

—— 著 者 略 歴 ——

1984 年　東京工業大学工学部情報工学科卒業
1989 年　東京工業大学大学院博士課程修了（情報工学専攻）
　　工学博士
1989 年　東京工業大学助手
1992 年　北陸先端科学技術大学院大学助教授
2000 年　東京工業大学助教授
2007 年　東京工業大学准教授
2009 年　東京工業大学教授
　　現在に至る

自然言語処理の基礎

Introduction to Natural Language Processing

2010 年 10 月 28 日　初版第 1 刷発行

検印省略

著　者　奥村（おくむら）　学（まなぶ）
発 行 者　株式会社　コロナ社
　　代 表 者　牛来真也
印 刷 所　三美印刷株式会社

112–0011　東京都文京区千石 4–46–10
発行所　株式会社　コロナ社
CORONA PUBLISHING CO., LTD.
Tokyo Japan
振替 00140–8–14844・電話(03)3941–3131(代)
ホームページ http://www.coronasha.co.jp

ISBN 978–4–339–02451–7　（金）　（製本：牧製本印刷）
Printed in Japan