自然言語処理概論
―組合せ最適化の観点から―

西川　仁

自然言語処理は人間が日常的に用いる自然言語を計算機を用いて処理する技術である．自然言語，特にテキストとして記述された言語は，離散的な記号，すなわち文字の列として表現されるため，その処理のために広く組合せ最適化の知見が取り入れられてきた．自然言語処理は大きく自然言語解析と自然言語生成の 2 種類に分けることができ，これらに含まれる個々の技術の詳細についてはそれぞれ本特集内の個別の記事に譲るとして，本稿では自然言語処理の全体的な枠組みについて組合せ最適化の観点から解説する．

キーワード：自然言語処理，組合せ最適化，数理最適化

1.　自然言語処理と探索

まず，以下の，20 文字からなる日本語の文を考える．

(1) 発達した低気圧のため東京は大雨となった．

一般に，日本語で書かれた文は文を構成する語の境界が明確でない，すなわち分かち書きされていないため，この文を機械に解釈させようとする場合，まず語の境界を明らかにする必要がある[1]．以下は形態素解析器 JUMAN[2] による分かち書き[3] の結果である．「｜」は語境界を示す．

(2) 発達｜した｜低｜気圧｜の｜ため｜東京｜は｜大雨｜と｜なった｜．

この例には 11 箇所の語境界が存在するが，20 文字からなる文には，潜在的には，19 箇所の語境界が存在する．それぞれの箇所について，当該箇所が語境界であるか，そうでないかの 2 通りが考えられるため，潜在的な分割候補は 2 の 19 乗にもなる．もちろん，候補の中には，すべての語境界候補を語境界でないと考え 20 文字すべてを 1 語とするものや，1 文字を 1 語としてすべての語境界候補を語境界として 20 語からなる文とするものなど，明らかに誤りだと考えられるものが無数に存在する．以下，例を示す．

(3) 発達した低気圧のため東京は大雨となった．

(4) 発｜達｜し｜た｜低｜気｜圧｜の｜た｜め｜東｜京｜は｜大｜雨｜と｜な｜っ｜た｜．

(3) はこの 20 文字からなる文を一つの単語とみなした場合，(4) はすべての文字それぞれを 1 単語とみなした場合である．ほかにも，(2) の代わりに，たとえば「低｜気圧」ではなく「低気圧」と 1 語としたり，また「大雨」を「大｜雨」とするものも考えられよう．これらの潜在的な分割候補は無数に存在するため，これらの中から，何らかの基準に従って，よい分割候補だと思われるものを効率的に探索するという要請が生じる．

次に，機械に文を生成させることを考える．非常に単純に考えると，10 単語の文を生成する際に，計算機が利用可能な語彙に 5 万の語が含まれているとすると，生成可能な文の種類は 5 万の 10 乗となる．しかし，もちろん，この莫大な，潜在的に生成可能な文のほぼすべては何ら意味をなさない，まったく無意味な単語列である．たとえば，格助詞の「を」が 10 回繰り返されているだけの文は（少なくとも多くの読み手にとっては）何ら意味をなさない単なる文字列である．以下は，本節のテキストを形態素解析器 JUMAN に入力し，その出力からランダムに単語を並べた例である．

(5) 探索｜まま｜学｜規模｜参考｜候補｜の｜ため｜ない｜多く

(6) これ｜(｜連｜対象｜素早く｜存在｜目的｜知見｜最大｜もの

言うまでもなく，(5) と (6) のいずれも，文というよりはランダムな単語の列に過ぎない．そのため，少なくとも，文法的であり，意味をなし，かつ，この文

にしかわ　ひとし
東京工業大学情報理工学院
〒 152–8550　東京都目黒区大岡山 2–12–1 W8–73
hitoshi@c.titech.ac.jp

[1] もちろん，単にキーワードを検出するだけであれば，テキストを走査し事前に用意されたキーワードと照合するだけでよい．しかし，たとえば「東京都」という文字列には「京都」が含まれており，地方自治体としての京都府あるいは京都市に関するテキストを探そうとした際に誤って「東京都」という文字列を含むテキストが検索されてしまうことは問題であろう．この問題を解決するためには分かち書きが必要となる．

[2] http://nlp.ist.i.kyoto-u.ac.jp/?JUMAN

[3] 形態素解析については 2 節で述べる．

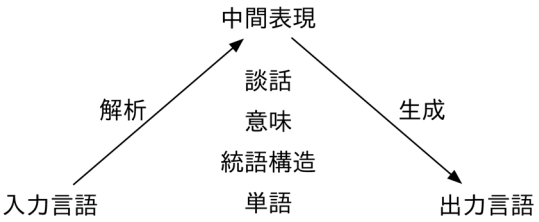

図 1　自然言語処理における解析と生成

が生成される目的，たとえばユーザによって入力された特定の質問に対する返答となっているなどの基準を満たした文のみを，この潜在的に生成可能な文の集合から効率的に探索するという要請が生じる．

このように，自然言語処理は離散的な記号列を対象とする処理であるため，何らかの処理を行う際には，無数の組合せの中から何らかの基準に沿って，良好な性質を保持し，多くの場合においては最も良好なものを素早く探し出すことが必要となる．これを実現するため，自然言語処理の分野では組合せ最適化の分野からさまざまな知見を導入しており，特に計算機の価格が相対的に低下し高速な計算資源を安価に利用することが可能になって以来，整数計画法など組合せ最適化分野で研究されてきた手法が広く利用されるようになっている．

再び機械に文を生成させることを考える．たとえば，日本語の書き言葉においては，格助詞「を」が 2 回連続して現れることは，入力の誤りの場合を除いてまず見られないであろう[4]から，大規模なテキストの集合[5]から単語の連接に関する統計量を取ることによって，日本語らしい文をある程度定量的に把握することができるであろう．このような何らかの統計量あるいは機械学習によって得られる解候補の解としての良好さを求める関数，すなわち目的関数を得ることができれば，あとはこの目的関数を最大とする解を探索することが目的となる．このように，自然言語処理において探索は決定的に重要な役割を果たしている．

自然言語処理は，伝統的に，単語のような「浅い」解析から意味や文脈といった「深い」解析へと段階的に解析を深め，究極的には，述語論理のような，特定の自然言語に依存しない何らかの中間表現を得るというアプローチを取っている（図 1）．もちろんアプリケーションによって必要となる解析の深さは異なるため，すべての自然言語処理アプリケーションが図 1 に示すような手続きを取るわけではない．また近年では End-to-End アプローチと呼ばれる，中間的な解析を陽に行わず，入力から直接目標となる出力を得るアプローチも盛んに研究されている．一方で，依然として解析の各段階は個別の課題として研究されてもいるため，自然言語の解析と生成それぞれにおける組合せ最適化の利用の詳細については個々の記事に譲り，まず自然言語処理の全体像について概説する．その際，関連のある文献をあわせて紹介する．より詳しい内容について関心のある読者は，個別の記事とともに参考文献を参照されたい．自然言語処理分野全体を見渡すための文献として，言語処理学会による言語処理学事典 [1]，人工知能学会による人工知能学大事典 [2] の自然言語処理の章がある．

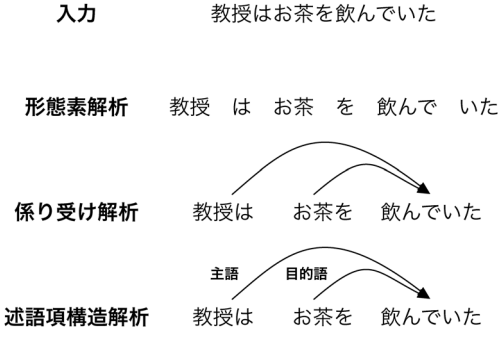

図 2　自然言語解析の手続き

2.　自然言語の解析

自然言語解析は何らかの文や文章を入力として受け取り，その入力を構成する単語や単語同士の統語構造，それらから読み取ることのできる意味を抽出する処理である．大まかな手続きの例を図 2 に示す．形態素解析や係り受け解析については後述するが，「教授はお茶を飲んでいた」というような入力が与えられた際に，段階的に，異なる複数の解析を入力に対して行っていくアプローチが自然言語解析の伝統的な方針である．

[4] もちろん，今日ではチャットなどにおいて，たとえば驚きの表現として「ををを…」などといった発言がなされることはままある．このようにチャットなどが広く普及する前にはあまり目にかかることのなかった表現への対処も自然言語処理の重要な課題である．

[5] これを「コーパス」と呼ぶ．

機械翻訳をアプリケーションとして考えると，例に示すように，この文の述語（動詞）が「飲んでいた」であり，主語が「教授」，目的語が「お茶」であるとわかれば，ごくごく単純に考えると，これらの表現をそれぞれ訳し，また英語の語順に当てはめることによって“The professor was drinking tea”という訳文が得られる．このように，自然言語解析の結果を適切に利用することによって，さまざまなアプリケーションが可能になる．

以下，自然言語解析に含まれる典型的な処理を個別に概説する．

2.1 形態素解析

一般に，日本語のテキストは分かち書きされない．そのため，計算機において自然言語を処理する最初の処理は形態素解析と呼ばれ，文を構成する語の境界を明らかにし，またそれぞれの語の，名詞や形容詞といった品詞を明らかにする．

典型的には，形態素解析は，まず入力された文の頭から辞書を参照しつつラティスと呼ばれる有向非循環グラフを生成し，そのうえでグラフを構成する各ノードの出現しやすさ，すなわちそれぞれの語が大規模なコーパスにおいて出現する頻度と，それぞれの語が隣接して出現する頻度などをもとに，最ももっともらしい語の列を選択する．探索はビタビ・アルゴリズムと呼ばれる動的計画法の一種で行うことができる．

2.2 係り受け解析

文をなす語の境界とその品詞が明らかになった後はその文の統語構造を明らかにする処理が行われる．統語構造は文を構成する単語同士の関係がなす構造である．日本語では統語解析として文を構成する語同士の係り受け関係を明らかにする，係り受け解析と呼ばれる処理が行われる．係り受け解析は自然言語解析において最も盛んに研究が行われている分野であり，また整数計画問題としての定式化が盛んな分野でもある．詳しくは本特集号の「自然言語解析―整数計画問題としての定式化と解法―」を参照されたい．

2.3 意味解析

文の統語構造が明らかにされた後，文の意味的な構造を得る処理が行われる．この処理は述語項構造解析と呼ばれる．たとえば，「二郎はラーメンを食べた」という文を考えると，この文の主語は「二郎」であり，「ラーメン」が目的語であり，「食べた」が述語となる．また，日本語のテキストにおいては主語や目的語は頻繁に省略され，「それ」や「これ」といった指示代名詞が利用されることもある．このようにテキストに現れる述語とその主語，目的語の関係は複雑であり，複数の述語の整合性を同時に考慮する必要性が生じるため，整数計画法を用いて複数の関係を同時に推定する取り組みもある [3]．述語項構造解析を扱う書籍としては笹野と飯田によるもの [4] が詳しい．

2.4 談話解析

最後に，文をまたぐ関係の解析が行われる．たとえば，「二郎は朝食を食べ損ねた．そのため，昼食は多めに取った」という文を考えると，最初の「二郎は朝食を食べ損ねた」という文は，次の「そのため，昼食は多めに取った」という文の原因となっていることがわかる．このような解析は文章全体の論理的な構造を分析するために必要であり，文章全体を処理の対象とするアプリケーション，たとえば自動要約などにとって重要である．

3. 自然言語の生成

自然言語生成は，広義には，自然言語処理において，機械に何らかのテキストを生成させる処理全般を指す．その中には機械翻訳，対話処理，質問応答，自動要約など，今日においてはすでに身近となっている複数のアプリケーションが含まれる．

狭義には，何らかの中間表現，たとえば天気予報や株価の変動など，明に表現すべき事柄が定まっているある種のデータを入力とし，それを自然なテキストとして表現する処理である．

これらの処理においては，必ずしも単語を一から組合せてテキストを生成する必要はなく，事前に用意したテンプレートに適切な情報を挿入する，既存のテキストの一部を組合せて新しいテキストを生成する，また質問に対する回答としてふさわしい文などを単に検索しそれを回答とするなど，さまざまな方法で対処することができる．

3.1 機械翻訳

機械翻訳は自然言語処理最大のアプリケーションであり，自然言語処理の歴史はそのまま機械翻訳の歴史といっても過言ではない．機械翻訳は，典型的にはある言語で書かれた単一の文を，異なる言語において意味的に同一の文に変換する処理である．機械翻訳における訳出は莫大な単語あるいは単語の複合によって構成される句の組合せの中から，訳文として適切なものを選択する処理であり，効率的な探索が本質的に重要であるため，統計的な手法による機械翻訳が主流となって以来，整数計画法が盛んに利用されている．機械翻訳全般を扱った書籍として渡辺らによるもの [5] がある．また

特に組合せ最適化に焦点を絞ったものとして Neubig and Watanabe によるサーベイ論文 [6] がある．

3.2 対話システム

機械翻訳と並び，自然言語処理の代表的なアプリケーションが対話システムである．対話処理は機械翻訳と同様に何らかの入力文，典型的にはユーザの発話あるいは入力を受け取り，それに対応する適切な応答を生成する．応答の生成に際しては入力されたテキストに含意される情報のみならず，機械に向かっているユーザの性質や状態も考慮する必要があるため，洗練された対話システムはさまざまな異なる要素技術が連携して動作する．対話処理を網羅的に扱った書籍としては中野らによるもの [7] がある．

3.3 自動要約

自動要約は，その最も単純な形態としては，与えられた単一の文書を入力として，文字数や文数などで定義される所定の長さ以内の要約を生成する課題である．入力が複数の文書である場合など，さまざまな派生課題が存在するものの，典型的には入力文書中に含まれる重要な情報を特定したうえで，その情報が要約に含まれるように入力文書に含まれる文を抽出，あるいは新しく文を生成し要約を構成する．要約の長さという制約下において可能な限り情報を要約に詰め込むという課題の性質上，自動要約はナップサック問題としての定式化が可能であり，組合せ最適化と極めて親和的な課題である．このことから，組合せ最適化分野の知見が盛んに導入されている [8]．詳しくは本特集号の「文書要約のための数理的手法」を参照されたい．

3.4 質問応答

クイズ番組 Jeopardy! において人間を破った IBM 社による Watson，また Siri や Cortana のようなアプリケーションなど，質問応答システムはすでに身近に使われている．質問応答システムは，一般的には，雑談などとは異なり，明確な質問の意図をもつ問いかけに対して適切な回答を返すものとして設計される．質問には，「富士山の高さ」のような数値や「スペインの首都」のような地名など，ある種のデータベースを探索することによって回答が可能なものから，「空はなぜ青い」といったある種の説明を生成する必要があるものまで含まれる．

後者のような質問については，大規模なテキスト集合に対する適切な検索および関連するテキストからの回答となる部位の適切な抽出が必要となることから，情報検索および自動要約が内部で利用されることもある．

4. その他のアプリケーション

上で述べた形態素解析などの要素技術単体でも十分有用なものであるが，特定の用途のために研究が行われている自然言語処理課題として以下のようなものがある．

4.1 固有表現抽出

固有表現抽出は，地名や人名，企業名，製品名といった固有表現や，日付などの数値表現をテキストの中から抽出する処理である．以下の文を考える．

(7) 田中角栄は，1972 年から 1974 年にかけて日本の総理大臣を務めた．

この文からは，「田中角栄」という人名，「日本」という地名，「1972 年」および「1974 年」という数値表現を抽出することができる．これらの情報は自然言語処理の商業的な応用において特に重要である．たとえば，テキストから企業名や製品名を抽出することができれば，後で述べる評判分析とあわせ，ある企業のある製品名が好ましくない文脈で頻繁に出現しているといったことを調査することができる．このような応用がマーケティングにおいて有用であることは言を俟たないであろう．技術的には，固有表現抽出は形態素解析と同様にタグ付け課題として定式化することができる．日本語のテキストに対してこれを行う場合，形態素解析がすでに行われたテキストに対して，固有表現を構成する一連なりの語にそれらの語がある固有表現を形成していることを示すタグが付与される．その後，タグが付与された大規模な学習データを用意したうえで，機械学習を利用して抽出器が訓練される．

4.2 関係抽出

関係抽出は，固有表現間のある種の関係を抽出する課題である．以下のような文を例として考える．

(8) 富士山の標高は 3,776 メートルである．

この文には，「富士山」という地名と「3,776 メートル」という数値表現が含まれており，この文からはこの二つの間に「標高」という関係があることがわかる．そのため，この文からは，「富士山，3,776 メートル，標高」という三つ組を抽出することができる．大規模なコーパスに対して関係抽出を実行することによって，このようなさまざまな知識をテキストから獲得することができ，獲得された知識は質問応答などに利用することができる．

以降，ある文に，二つの固有表現候補 e_1 と e_2 が存在しており，ある固有表現抽出器が出力する，それぞれが固有表現である確率を p_{e_1} と p_{e_2} とする．また，

表 1　固有表現抽出と関係抽出

パタン	x_{e_1}	x_{e_2}	x_r
1	0	0	0
2	1	0	0
3	0	1	0
4	1	1	0
5	1	1	1

これら二つの固有表現候補に対して，関係抽出器がこれらの間に何らかの関係 r があるとする確率を p_r とする．抽出される三つ組は，$\langle e_1, e_2, r\rangle$ となる．

固有表現間の関係を抽出するためには，まず固有表現抽出を実施し，その後，抽出された二つの固有表現の間の関係の有無を判定することになる．そのため，e_1 と e_2 が実際に固有表現であると固有表現抽出器に判定されたときのみ，関係 r の有無がそれに続いて判定されることになる．しかし，このとき，仮に，p_1 と p_2 が低い値であったとしても，後段の関係抽出器が e_1 と e_2 の間に極めて高い確率で何らかの関係が存在すると判定しうるとすると，このことを手がかりに，e_1 と e_2 を固有表現として判定するべきかもしれない．このような手がかりを利用するためには，関係抽出を固有表現抽出の後に行うのではなく，固有表現抽出と同時に行う必要がある．

ここで，x_{e_1} と x_{e_2} をそれぞれ，e_1 と e_2 がそれぞれ固有表現であると判定されるときに 1，そうでないときに 0 を取る変数とする．また，x_r を関係 r が存在すると判定されるときに 1，そうでないときに 0 を取る変数とする．これら三つの変数が取りうる値の組み合わせは 2^3 通りとなるが，x_{e_1} と x_{e_2} がともに 1 であるときのみ x_r は 1 を取りうるため，可能な組合せは 5 通りとなる（表 1）．

このとき，$s_i = \log \frac{p_i}{1-p_i}$ として，以下のような整数計画問題を解くことによって，固有表現抽出と関係抽出を同時に行うことができる．

$$\begin{aligned} \max_{x_{e_1}, x_{e_2}, x_r} \quad & \{s_{e_1} x_{e_1} + s_{e_2} x_{e_2} + s_r x_r\} \\ \text{s.t.} \quad & x_{e_1}, x_{e_2}, x_r \in \{0, 1\} \\ & x_r \leq x_{e_1} \\ & x_r \leq x_{e_2} \end{aligned}$$

このように，前段の処理と後段の処理を同時に実施することで，精度の向上を図ることができる一方，課題としてはより複雑になり，前段の処理の解と後段の処理の解の組合せを考慮する必要が生じる．Roth and Yih は前段の処理となる固有表現抽出と後段の処理となる関係抽出を整数計画問題として定式化し，これらを同時に解くことによって，関係抽出の精度が向上したことを報告している [9]．このように，同時に二つ以上の異なる自然言語処理課題を同時に解決することによって精度の向上を図る際，整数計画問題としての定式化は強力な武器となる．

4.3　評判分析

ある特定の商品名や企業名が，好ましい文脈で出現しているのか，あるいは好ましくない文脈で出現しているのか判断するものである．単純な手法としては，ある特定の製品名の周辺に出現している単語の傾向を調査するものがある．たとえば，ツイートを大規模に収集したうえで，ある特定の洗濯用洗剤の周囲に好ましくない単語，たとえば「悪い」「汚い」などが頻繁に出現しているのであれば，当該商品を販売している企業は何らかの対策を取る必要があるであろう．文献としては大塚らによるもの [10]，Pang and Lee によるもの [11] がある．

4.4　テキスト含意認識

二つの言明が与えられたときに，一つがもう一つの言明を含意するか，それともそれらが矛盾しているか，あるいは全く無関係であるかを判定する処理である．たとえば以下のような言明を考える．

(9) ビタビ・アルゴリズムはアンドリュー・ビタビによって考え出された．
(10) アンドリュー・ビタビはビタビ・アルゴリズムの考案者である．
(11) 鈴木二郎が開発したアルゴリズムの一つにビタビ・アルゴリズムがある．
(12) 自然言語処理は日本語や英語などの自然言語で書かれたテキストを計算機を用いて処理する技術である．

言明 (9) が与えられたとすると，言明 (9) は言明 (10) を含意する．一方，言明 (9) と言明 (11) は矛盾している．また，言明 (9) と言明 (12) は無関係である．テキスト含意認識は自動要約や質問応答などのアプリケーション内部で利用されるほか，今日では真偽不明の報道がソーシャルメディア上で頻繁に飛び交うため，これらの報道に対する信憑性の判断への応用も期待されている．

5.　おわりに

本稿では自然言語処理について組合せ最適化の観点から概説した．自然言語解析と自然言語生成それぞれについての詳細は，本特集の個別の記事を参照されたい．

上で述べたように，自然言語処理は離散的な記号列を対象とするという性質から，潜在的な解候補集合の中から良好な解を素早く見つけ出す必要に迫られており，そのため組合せ最適化分野で開発されてきたさまざまな手法を導入してきた．今後は，深層学習に基づくモデルにおけるより洗練された探索が，自然言語処理における組合せ最適化の利用先として重要であろう．現在は，深層学習に基づくモデルにおいてはビームサーチなど比較的単純な方法で探索が行われているが，今後この部分を改良する方向に研究が進むことは，過去の自然言語処理研究の流れを鑑みると疑いない．

本稿が組合せ最適化分野と自然言語処理の橋渡しの一助となれば幸甚である．

謝辞　本稿の執筆に際しては，情報通信研究機構飯田龍主任研究員，専修大学高野祐一准教授，東京工業大学徳永健伸教授よりさまざまなご意見を頂戴した．記して感謝する．

参考文献

[1] 言語処理学会（編），『言語処理学事典』，共立出版，2009.
[2] 人工知能学会（編），『人工知能学大事典』．共立出版，2017.
[3] R. Iida and M. Poesio, “A cross-lingual ilp solution to zero anaphora resolution,” In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011)*, pp. 804–813, 2011.
[4] 笹野遼平，飯田龍，『文脈解析―述語項構造・照応・談話構造の解析―』，コロナ社，2017.
[5] 渡辺太郎，今村賢治，賀沢秀人，G. Neubig, 中澤敏明，『機械翻訳』，コロナ社，2014.
[6] G. Neubig and T. Watanabe, “Optimization for statistical machine translation: A survey,” *Computational Linguistics*, **42**(1), pp. 1–54, 2016.
[7] 中野幹生，駒谷和範，船越孝太郎，中野有紀子，『対話システム』，コロナ社，2015.
[8] A. Nenkova and K. McKeown, *Automatic Summarization*, Now Publishers, 2011.
[9] D. Roth and W.-T. Yih, “A linear programming formulation for global inference in natural language tasks,” In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pp. 1–8, 2004.
[10] 大塚裕子，乾孝司，奥村学，『意見分析エンジン―計算言語学と社会学の接点―』，コロナ社，2007.
[11] B. Pang and L. Lee, *Opinion Mining and Sentiment Analysis*, Now Publishers, 2008.

