ま　え　が　き

翻訳は，専門的な訓練を受けた翻訳者あるいは通訳者により，何世紀もの間行われてきた知的な作業といえる。その作業を機械により置き換える，あるいは，機械の支援を受けることで大幅にその作業効率がよくなるであろう。機械翻訳は，コンピュータの黎明期から，アプリケーションの一つとして長年研究開発されてきた技術である。初期は言語学的な知識を記述することで実現されていたが，コンピュータの発達およびコーパスなどの言語データが整備されるに伴い，データから自動的に知識を抽出する手法が主流になってきた。本書は，特に最近のトピックである，機械学習および統計的手法に基づいた，統計的機械翻訳を中心に解説する。

統計的機械翻訳は，複雑な統計モデルに基づいたさまざまな要素技術により構成されているが，2000 年代に入ってからツールが整備された結果，オンライン翻訳サービスや携帯端末での音声翻訳アプリ，ビジネス向けのツールとして実用化されている。研究開発の速度は凄まじく，関連した研究も含めると，ほぼ毎年 200～300 件の論文が出版されている。このため，大学生や大学院生，また，言語処理の他の分野の研究者でさえ，覚えることが多く，機械翻訳を研究課題として新たに始めるのは，非常にハードルが高い。

本書は，各要素技術が「何をしているのか」だけでなく，「どのように実現しているのか」その基礎的な知識まで体系的かつ詳細に説明することを心がけた。この結果，機械学習だけでなく，形式言語や構文解析，探索技術など，計算言語学の基礎技術を網羅することになったが，あくまでも機械翻訳の視点から解説したものであり，各分野の専門家には物足りないものとなったかもしれない。ただ，本書をきっかけとして，機械翻訳への理解が深まり，各種ツールをブラックボックスとして使う立場でなく，ツールをつくり，さらに改良を加える立場

になることを期待したい。また，今後さらに研究開発が進み，高精度な翻訳を可能とするシステムの実現へとつながれば幸いである。

知識に基づく機械翻訳および用例に基づく機械翻訳はそれぞれ今村，中澤が執筆した。機械翻訳の評価は主に賀沢が担当し，今村，Neubig が加筆した。言語モデルおよび章末問題，付録は Neubig が執筆し，残りの章は渡辺が担当した。東京工業大学の奥村学教授には，本書のご提案および監修をいただき，非常に感謝している。コロナ社の皆様には，企画から原稿の編集作業まで，出版に向けてたいへんなご尽力をいただいた。NTT コミュニケーション科学基礎研究所の林克彦さんには，さまざまな改善点や誤りを指摘していただき，感謝している。奈良先端科学技術大学院大学の学生さんたちにも貴重なご意見をいただいた。本書の内容は，以前行ったさまざまな講義あるいはチュートリアル講演などの資料を基にした。このような講演の機会を与えてくださった，京都大学の河原達也教授，ハルビン工科大学の Tiejun Zhao 教授，ならびに言語処理学会実行委員，高度言語情報融合フォーラムの委員の方々に感謝したい。

2013 年 12 月

渡辺 太郎
今村 賢治
賀沢 秀人
Neubig, Graham
中澤 敏明

目　　次

1. 機 械 翻 訳

2. 機械翻訳の評価

3. 言 語 モ デ ル

4. 単語アライメント

5. 句に基づく機械翻訳

6. 木構造に基づく機械翻訳

7. 最 適 化

1 機械翻訳

> ロシア語の文章を見て，思ったのは「これは本当は英語で書かれているが，変なシンボルで暗号化されている。ではいまから解読しよう。」

機械翻訳は，1947 年 3 月 4 日，Warren Weaver が Nobert Wiener に送ったこの手紙の内容[236] に集約されている。翻訳の問題を暗号解読として捉える考え方に対し，N. Wiener は，1947 年 4 月 30 日の返信で非常に悲観的な意見を示している[236]。

> 機械的な翻訳の問題については，異なる言語間の単語の境界は曖昧すぎて，さらに，感情や暗示的な意味などを考えると，擬似機械的な翻訳の枠組みでも捉えられないのではないか，と懸念している。

W. Weaver は 1947 年 5 月 9 日の手紙[236] で，楽観的な意見を唱えている。

> 例えば，2 000 単語と仮定して，二単語の組合せを一単語とみなしてすべての組合せを考えたとしよう。語彙数はたかだか 400 万単語で，現代のコンピュータでは大した数字ではないのでは。

機械翻訳は，この楽観的，悲観的という両極端な考え方を織り込みつつ，コンピュータの黎明期から重要なアプリケーションの一つとして研究開発が進められてきた。その歴史を 1.1 節で簡単に振り返り，代表的な機械翻訳の手法である，知識に基づく機械翻訳（1.2 節），用例に基づく機械翻訳（1.3 節）および統計的機械翻訳（1.4 節）について解説する。

1.1 歴　　史

1.1.1 初期の機械翻訳

W. Weaver と N. Wiener との手紙のやり取りの 2 年後，W. Weaver は，Andrew Booth と議論し，1949 年 7 月に機械翻訳のアイデアを覚書として書いている[236]。この覚書では，情報理論[200] に基づいた暗号解読，統計的手法だけではなく言語学的な知見を取り入れ，その可能性について論じており，これを契機に大学，研究所などで機械翻訳の研究が盛んになった。初期の機械翻訳研究の成果は，1954 年 1 月 7 日，Georgetown 大学と IBM が共同で開発した，ロシア語から英語への機械翻訳デモンストレーションとして発表された[108]。

機械翻訳システムの研究開発は，試行錯誤が繰り返されたが，機械翻訳に必要な基礎技術の研究開発を促し，**計算言語学**（computational linguistics）と呼ばれる，計算機による自然言語の処理という新しい分野が誕生した。**知識に基づく機械翻訳**（knowledge-based machine translation，**KBMT**）は，このような計算言語学の基礎技術を組み合わせ，原言語の解析，中間表現の変換，目的言語の生成から構成され，各段階で用いられる処理の粒度で分類される（1.2 節 参照）。

1.1.2 ALPAC レポート

全自動高品質機械翻訳への期待が高まる中，研究開発は遅々として進まず，この状況を分析するため，NSF（National Science Foundation）は 1964 年に ALPAC（Automatic Language Processing Advisory Committee）を組織した[108]。人間による翻訳や，機械翻訳を実現するための専門家の育成などあらゆる角度から費用対効果を分析した，いわゆる **ALPAC レポート**[52] では，「役に立つ機械翻訳は今も，また，今後も期待できない。」とし，機械翻訳への投資に対して疑問を投げかけた。さらに，言語学を科学の一分野とした計算言語学の各要素技術の研究開発および翻訳者への支援ツールの整備を勧告している。

ALPAC レポートの結果，アメリカでは機械翻訳のプロジェクトに対して予算がつかず，10 年以上研究開発が停滞した。

1.1.3 実用システム

ALPAC レポートでは「完璧な翻訳」という非常に高い目標を想定していたが，例えば，翻訳の目的が情報収集および**情報理解**（assimilation）である場合，ある程度低い精度であっても問題ないことが多い。また，翻訳者による文書も出版されるまでに何度も校正，編集されることを完全に無視していた。Systran は，最初の商用システムの一つであり，1970 年に露英機械翻訳がアメリカ空軍へ納入され，情報収集および分析目的で活用されている[207]。また，1976 年には Systran の英仏機械翻訳が EC 委員会（現在の欧州委員会）へ納入され，その後さまざまな言語対へと拡張されている。アメリカ空軍と異なり，**情報拡散**（dissemination）が目的であり，機械翻訳の結果を**後編集**（post-editing）することで対処している。

初期の完全に自動的な機械翻訳システムの代表例は MÉTÉO であり，モントリオール大学の TAUM プロジェクトで開発された英仏および仏英翻訳システムを基にしている[207]。MÉTÉO はそのドメインを気象情報へと制限することで高精度な機械翻訳を実現し，1976 年より運用を開始している。

80 年代になると Logos や METAL などが商用システムとして実現されている。1982 年には，日本でも科学技術庁（現在の文部科学省）が主導した Mu プロジェクト[166),226)] が開始され，科学技術文献の日英英日翻訳システムとして実現された。このプロジェクトの成果は，富士通や日本電気，東芝，日立，沖電気など日本における機械翻訳システムの研究開発に大きな影響を与えた[108]。80 年代には他にも NTT の ALT-J/E システムなど翻訳システムが研究され，そこで開発された辞書などのリソースが，機械翻訳以外の自然言語処理に転用された[109]。

1.1.4 データに基づく手法

1980 年代までの主要な機械翻訳の研究開発は言語学的な知見を基にしていたが，対訳データなど，言語リソースが充実するのに伴い，データに基づく手法が徐々に取り入れられるようになる。1981 年には，アナロジーによる機械翻訳が提唱され，機械翻訳は，原言語の入力文に対して近い訳文を検索し，その例文に基づき編集を行う，といった過程で表現された[165]。これは 1.3 節の**用例に基づく機械翻訳**（example-based machine translation，**EBMT**）として実現されているが，実際の研究開発は 1980 年代後半から始まっている。

1980 年代後半には，IBM の研究グループ†が 1940 年代の暗号解読の考え方をそのまま復活し[21]，Candide システム[9] として最初の**統計的機械翻訳**（statistical machine translation，**SMT**）を実現した（1.4 節 参照）。機械翻訳は，1999 年にジョンズ・ホプキンズ大学のワークショップで IBM の研究成果を再実装し，最初のツールキットが使われるようになってから急激に研究開発が進み，DARPA によって TIDES や GALE などの機械翻訳を中心としたプロジェクトが主導された。この間に，統計的手法による単語アライメント（4 章 参照）および**句に基づく機械翻訳**（phase-based machine translation，**PBMT**，5 章 参照）が実現され，2006 年に開始された Google 翻訳や Microsoft による Bing などのオンライン翻訳サービス，あるいは SDL によるビジネス向けの翻訳アプリケーションなどに応用されている。また，ATR や Verbmobil など機械翻訳に音声認識技術を組み合わせた音声翻訳の可能性も同時に追求され，例えば 2010 年には，VoiceTra など，携帯端末向けのアプリケーションなどが NICT により実現されている。2000 年中ごろからの急激な研究開発の進展は，特に翻訳の自動評価技術（2 章 参照）によるところが大きく，NIST や WMT，IWSLT，NTCIR などの評価型ワークショップで同条件で各手法を比較し，研究者が直接議論してきた。また，機械翻訳のさらなる性能向上へ向けていま現在も研究開発が続けられている。

† このときの研究者はほとんど金融業界へと移っていった。

1.2 知識に基づく機械翻訳

機械翻訳の歴史は，ほぼそのまま，知識に基づく機械翻訳の歴史ともいえ，長年かけた研究開発においてさまざまな手法が提案され，いまでも主要な機械翻訳方式として使用されている。まず，翻訳の難しさを議論し，代表的な手法について紹介する。

1.2.1 言語のずれと翻訳の難しさ

機械翻訳が難しい原因を，もし一言でまとめると，同じ意図（意味）を伝えようとしている文にもかかわらず，言語が異なると異なる表出をしなければならず，現象が1対1に対応しないためである。つまり，異なる言語の表出形同士には，さまざまなずれがある。英語と日本語の翻訳を考えた場合，以下のようなずれが存在する。

〔**1**〕 **語彙的なずれ**　単語の訳は一般的には複数あり，1対1に対応しない。英語と日本語間の例を挙げると，以下のものがある。

（**a**） **多義語**　単語には，同じ語でも複数の意味をもつ，多義語が存在する。英単語 “bank” が最も有名である。“bank” には「銀行」という意味と「土手」という，最低限二つの語義が存在する。日本語にしたときに，どちらの訳を選択するかは，英語の**語義曖昧性解消**（word sense disambiguation）の問題である。

（**b**） **粒度の違い**　たとえ語義が決定できたとしても，原言語と目的言語で単語の粒度が異なる場合がある。例えば，“rice” の日本語訳は，少なくとも「米」「ご飯」「稲」の三つが存在する。このうち，「米」と「稲」の意味は，英語のシソーラスの一種である **WordNet**[76] にも異なる語義として登録されているが，「ご飯」は「米」と同義として扱われている。このように，文化的背景などの違いから，単語の意味する範囲が異なる場合があり，語義レベルでも1対1に対応しない。

粒度の違いは，動詞でも起こる。例えば，日本語の動詞「飛ぶ」には，「飛行する」と「飛び跳ねる」の意味がある。「飛行する」なら “fly”，「飛び跳ねる」なら “jump” にしなければならない。

これらの違いは，実は目的言語から見た原言語の語義曖昧性解消を必要としていることと等しい。だから，語彙的な訳し分けを行うためには，語義曖昧性解消で使われていた格フレーム（p.15 参照）などの利用が有効となる。

動詞に関しては，語義的な違いだけでなく，時制やアスペクトなど，動詞に付随する属性の違いを含む場合がある。例えば，“I know him.” の訳は「私は彼を知っている」が普通であろう。“know” の訳が「知る」ではなく「知っている」になるのは，英語の “know” 自体が継続的な意味を含んでいるのに対して，日本語の「知る」は比較的短時間の事象しか表さないためである。

〔2〕 局所構造のずれ　語彙が英語と日本語で一致していても，品詞が異なる場合がある。この問題は少々わかりづらいが，品詞が変わると構文構造も変化することが問題になる。適切な機能語を補って品詞変化がないような句にしないと，構文的につながらなくなる。

例えば，日本語にも英語にも形容詞という品詞が存在するので，“beautiful” の訳は「美しい」にしておけば，ほとんどの場合，問題ない。一方，値の正負を表す形容詞 “positive” は，日本語の形容詞に適当な訳がない。名詞「正」を訳語とした場合，“positive” が述語なら助動詞「だ」を補って，日本語でも述語にする必要がある。

英: The value is positive.

日: 値は正だ。

形容詞が名詞を修飾する要素として使われた場合は，助詞「の」を補って，連体修飾句を構成させなければならない。

英: the positive value

日: 正の値

〔3〕 大域構造のずれ　英語を勉強し出すとすぐに気が付くことに，日本語と英語では主語と目的語の位置が異なるというものがある。英語では主語

(subject)，動詞（verb)，目的語（object）という順序で記述されるが，日本語では主語，目的語，動詞という順序で記述されることが多い。したがって，単語を正しく翻訳するだけでなく，語順も変換しないと正しい翻訳にはならない。

さらに，日本語では，主語や目的語は格助詞を用いて区別するが，英語では動詞からの相対位置で両者を区別する。つまり，英日翻訳では，英語の相対位置を格助詞に変換する必要がある。

以上は文に動詞が一つしかない単文の話であるが，1 文の中には英語の関係節，日本語の埋込み文など，複数の動詞が含まれることもある。このような複雑な構造をもつ文を翻訳するためには，単文の接続や位置についても変換しなければならない。

〔4〕 各言語固有の特徴 各言語固有の現象には，原言語・目的言語間で対応する現象がないものもある。目的言語に必要な現象が原言語にない場合は，適切に補わなければならない。日本語と英語では，例えば以下の現象がある。

(1) 日本語には，“the”，“a” のような冠詞が存在しないため，日英翻訳を行う場合は，冠詞を適切に補わなければならない。また，英語は単数，複数を明確に区別するが，日本語では明示されることは少ないため，日英翻訳では単複を推測する必要がある。

(2) 逆に，英語には助数詞がなく，英日翻訳を行う際にはこれを適切に補わなければならない。助数詞とは，「3 人」「2 羽」のように数詞に後続する接尾辞のことである。

(3) 日本語では，主語の省略は比較的頻発する。英語では主語が必須であるため，日英翻訳の際には主語を補う必要がある。

日: 遊園地に行きたい。

英: **I** want to go to an amusement park.

(4) 英語の場合でも，実は翻訳対象が男性か女性かで訳を区別する場合がある。例えば，「鈴木さん」の訳は “Mr. Suzuki” か “Ms. Suzuki” かはあいまいである。

索　　引

【T】

【U】

【V】

【W】

—— 監修者・著者略歴 ——

奥村　学（おくむら　まなぶ）

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

渡辺　太郎（わたなべ　たろう）

1994年　京都大学工学部情報工学科卒業
1997年　京都大学大学院修士課程修了
（情報工学専攻）
2000年　Language and Information Technologies, School of Computer Science, Carnegie Mellon University, Master of Science 取得
2001年～05年　ATR 音声言語コミュニケーション研究所
2003年　京都大学大学院博士後期課程研究指導認定退学（知能情報学専攻）
2004年　博士（情報学）（京都大学）
2005年～08年　NTT コミュニケーション科学基礎研究所
2009年　独立行政法人情報通信研究機構
現在に至る

賀沢　秀人（かざわ　ひでと）

1993年　東京大学理学部物理学科卒業
1995年　東京大学大学院修士課程修了
（物理学専攻）
1995年　日本電信電話株式会社勤務
2006年　博士（工学）（奈良先端科学技術大学院大学）
2006年　グーグル株式会社勤務
現在に至る

中澤　敏明（なかざわ　としあき）

2005年　東京大学工学部電子情報工学科卒業
2007年　東京大学大学院修士課程修了
（情報理工学専攻）
2010年　京都大学大学院博士後期課程修了
博士(情報学)
2011年　京都大学特定助教
2013年　独立行政法人科学技術振興機構研究員
現在に至る

今村　賢治（いまむら　けんじ）

1985年　千葉大学工学部電気工学科卒業
1985年　日本電信電話株式会社勤務
1995年～98年　NTT ソフトウェア株式会社勤務
2000年～06年　株式会社国際電気通信基礎技術研究所
2004年　奈良先端科学技術大学院大学博士課程後期修了（情報処理学専攻）
博士（工学）
2006年　NTT サイバースペース研究所
（現 メディアインテリジェンス研究所）
現在に至る

Neubig, Graham（ニュービッグ　グラム）

2005年　イリノイ大学工学部コンピューターサイエンス学科卒業
2005年　兵庫県立但馬農業高等学校勤務
2006年～08年　兵庫県県庁勤務
2010年　京都大学大学院修士課程修了
（情報学専攻）
2012年　博士（情報学）（京都大学）
2012年　奈良先端科学技術大学院大学助教
現在に至る

機 械 翻 訳

Machine Translation

2014 年 2 月 21 日　初版第 1 刷発行

検印省略

監修者　奥村　学
著　者　渡辺太郎
　　今村賢治
　　賀沢秀人
　　Neubig, Graham
　　中澤敏明
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 978–4–339–02754–9　（金）（製本：愛千製本所）

Printed in Japan