

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

1.1 歴 史	2
1.1.1 初期の機械翻訳	2
1.1.2 ALPAC レポート	2
1.1.3 実用システム	3
1.1.4 データに基づく手法	4
1.2 知識に基づく機械翻訳	5
1.2.1 言語のずれと翻訳の難しさ	5
1.2.2 要素合成原理	8
1.2.3 翻訳のレベル	9
1.2.4 解 析	10
1.2.5 変 換	12
1.2.6 生 成	15
1.2.7 その他の話題	15
1.3 用例に基づく機械翻訳	16
1.3.1 事例ベース推論としての機械翻訳	18
1.3.2 用例の検索と修正	19
1.3.3 複数用例の利用	21
1.3.4 組合せの手がかり	23
1.3.5 最適な用例の組合せ	24
1.3.6 複雑な組合せ	25

1.3.7 今後の発展	26
1.4 統計的機械翻訳	27
1.4.1 暗号解読としての機械翻訳	27
1.4.2 モデル	29
1.4.3 学習	31
1.4.4 デコード	32
1.4.5 最適化	33
1.5 まとめ	35
章末問題	36

2. 機械翻訳の評価

2.1 機械翻訳を評価するとは	37
2.1.1 機械翻訳の用途	38
2.1.2 ブラックボックス評価・ガラスボックス評価	39
2.1.3 内的評価・外的評価	39
2.2 主観評価	40
2.2.1 グレード評価 (N 段階評価)	40
2.2.2 一対比較／ランキング	41
2.2.3 人間翻訳編集率	44
2.2.4 評価者について	45
2.2.5 評価の一致	45
2.3 自動評価	47
2.3.1 WER	48
2.3.2 TER	50
2.3.3 BLEU	51
2.3.4 METEOR	53

2.3.5	RIBES	55
2.3.6	自動評価のメタ評価	57
2.4	評価結果に基づく差分の検出	58
2.4.1	統計的に有意な差	58
2.4.2	評価値の信頼区間	58
2.4.3	対比較の有意差検定	59
2.5	まとめ	60
	章末問題	60

3. 言語モデル

3.1	n -gram モデルの基礎	63
3.2	n -gram モデルの平滑化	66
3.2.1	線形補間	66
3.2.2	Witten-Bell 法	67
3.2.3	絶対割引法	68
3.2.4	Kneser-Ney 法	70
3.2.5	その他の平滑化法	71
3.2.6	未定義語への対応	71
3.3	言語モデルの評価	72
3.3.1	尤度・対数尤度	72
3.3.2	エントロピー・パープレキシティ	73
3.3.3	カバレッジ	74
3.4	学習データと言語モデル性能	74
3.5	言語モデルの格納・参照	75
3.6	クラス n -gram モデル	76
3.7	まとめ	78

章 末 問 題	79
---------	----

4. 単語アライメント

4.1 ヒューリスティックモデル	81
4.2 IBM モデル	82
4.2.1 アライメントに基づくモデル	84
4.2.2 稔性に基づくモデル	91
4.2.3 学 習	99
4.3 両方向モデル	107
4.3.1 ヒューリスティック	108
4.3.2 事後確率によるフィルタリング	110
4.3.3 学習時の制約	111
4.4 教師あり単語アライメント	114
4.4.1 単語アライメントの評価	115
4.4.2 識別学習によるアライメント	117
4.5 ま と め	122
章 末 問 題	123

5. 句に基づく機械翻訳

5.1 句に基づく翻訳モデル	125
5.2 句に基づくモデルの学習	127
5.3 対数線形モデル	130
5.4 デ コ ー ダ	137
5.4.1 巡回セールスマン問題	138
5.4.2 動的計画法に基づくアルゴリズム	139

5.4.3	グラフ構造	142
5.4.4	半環	144
5.4.5	k -best の導出	147
5.4.6	探索空間の制約	148
5.4.7	ビーム探索	150
5.4.8	ヒューリスティック関数	152
5.5	フレーズペアの事前並び替え	155
5.5.1	事前並び替えルール	156
5.5.2	事前並び替えモデル	157
5.5.3	事前並び替え学習	158
5.6	まとめ	159
	章末問題	160

6. 木構造に基づく機械翻訳

6.1	文脈自由文法	162
6.1.1	構文解析	163
6.1.2	演繹システム	165
6.1.3	超グラフ	166
6.1.4	半環構文解析	169
6.1.5	k -best 導出	171
6.2	同期文脈自由文法	175
6.2.1	同期文脈自由文法の特徴	176
6.2.2	同期文脈自由文法の学習	179
6.2.3	統語論的ラベルの導入	182
6.2.4	素性	183
6.2.5	デコーディング	187

6.2.6	リ ス コ ア	188
6.3	同期木置換文法	193
6.3.1	同期木置換文法の特徴	195
6.3.2	同期木置換ルールの学習	196
6.3.3	素 性	204
6.3.4	デコーディング	206
6.3.5	二 分 化	211
6.4	二言語同期解析	214
6.4.1	反転トランスダクション文法	215
6.4.2	スパン枝刈り	217
6.4.3	ビーム探索	220
6.4.4	二段解析	223
6.5	ま と め	225
	章 末 問 題	226

7. 最 適 化

7.1	準 備	229
7.2	バ ッ チ 学 習	231
7.2.1	エラー最小化学習	232
7.2.2	確 率 モ デ ル	237
7.2.3	マージン最大化	239
7.2.4	ベイズリスク最小化	242
7.3	オンライン学習	247
7.3.1	エラー関数の近似	249
7.3.2	学習アルゴリズム	251
7.3.3	スパースな素性	255

7.3.4 並 列 化	258
7.4 ま と め	262
章 末 問 題	263
付録 機械翻訳のための資源・ツール	264
A.1 対 訳 デ ー タ	264
A.2 文 書 ・ 文 ア ラ イ メ ン ト	265
A.3 翻 訳 評 価	265
A.4 言 語 モ デ ル	266
A.5 単 語 ア ラ イ メ ン ト	266
A.6 機 械 翻 訳 シ ス テ ム	267
引 用 ・ 参 考 文 献	268
章 末 問 題 解 答	294
索 引	304

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、2000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てしよう。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年にジョーンズ・ホプキンス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切な訳がない。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

【あ】	演繹システム	165	確率的勾配降下法	118, 238, 254
ア－リー法	エンコード	29	確率モデル	117
アジェンダ	エントロピー	73	隠れ変数	30, 229
後編集	エントロピー枝刈り	130	隠れマルコフアライメント	
アナロジ－に基づく機械	エントロピー最大法	117	モデル	88
翻訳			下降法	172
アライメントエラーレート	【お】		加算平滑化	71
	重み	143	過剰適合	33
	重み付き非周期有向グラフ	142	括弧反転トランスダク	
アライメントスパン	重み付き非周期有向超	142	シオン文法	216
アライメントに基づく	グラフ	166	カップ係数	46
モデル	親方向の手がかり	24	カバレッジ	74
アライメントモデル	オラクル翻訳	238	ガラスボックス評価	39
	オンライン学習	228, 248	カルバック・ライブラー	
			ダイバージェンス	101
【い】	【か】		環	145
イェンゼンの不等式	開始頂点	143	簡潔ベナルティ	52
行き掛け順	解析処理	10		
位相的順序	階層のフレーズ	180	【き】	
依存構造解析	階層のフレーズ文法	180	木構造に基づく機械翻訳	
位置独立単語誤り率	階層のルール	180		207
一貫している	外的評価	39	擬似コーパス	249
一対比較	開発セット	35	基数	140
意味役割付与	開発データ	35, 67	木接合文法	24, 193
	改良型絶対割引法	70	期待損失	242
【う】	改良型 Kneser-Ney 法	71	期待値最大化アルゴリズム	100
後ろ向きアルゴリズム	帰りがけ順	167	期待 BLEU	243, 245
内側アルゴリズム	可換性	144	木置換文法	193
内側性能指数	可逆圧縮	76	基本木	193
内側外側アルゴリズム	学習	31	逆アライメント	91
	学習データ	64	逆アライメントモデル	92
【え】	学習率	254	キューブ枝刈り	190
枝刈り	確定的アニーリング	243	キューブプルーニング	190
エプシユタイン法	格フレーム	15		
エラー関数				
エラー最小化学習				

教師あり学習	32, 81, 114	語彙翻訳モデル	132	事前並び替え手法	155
教師なし学習	32, 81, 114	語彙モデル	85, 92	シソーラス	20
凝集クラスタリング	77	交換	134	始点	143, 167
共役勾配法	100, 238	交換アルゴリズム	78	シャード	258
局所的更新	238	後件	165	終端記号	162
局所的な素性	142, 188	交差	108, 163	終点	143, 166
許容可能な頂点	201	項数	167	終了頂点	143
近傍	96	合成	178	主観評価	40
近傍のアライメント	96	構造学習	35, 229	主辞駆動句構造文法	157
		構造変換	12	主辞後続型言語	157
【く】		構造ランプ損失	254	主辞交代	25
クーリング	243	勾配上昇法	100	主辞先行型言語	157
釘付けされたアライメント		構文解析	11, 163	述語項構造解析	12
	97	構文解析木	168	順位相関係数	55
句構造解析	11	構文解析森	168	巡回セールスマン問題	138
句に基づく機械翻訳	4, 125	構文拡張機械翻訳	182	準ニュートン法	100
句反転トランスダクション		構文トランスファー方式	10	条件付き確率場	118
文法	216	項目	165	上昇法	164
区分的線形	234	公理	165	状態	138
句翻訳モデル	127	コーパスに基づく機械翻訳		情報拡散	3, 38
組合せ範疇文法	182		18	情報交換	38
組合セルール	202	ゴール	165	情報理解	3, 38
句歪みモデル	127	語義曖昧性解消	5	将来スコア	153
クラスタリングアルゴリズム	77	子方向の手がかり	24	事例ベース推論	18
グレード評価	40	根	23	信頼区間	58
クロネッカーのデルタ	91				
		【さ】		【す】	
【け】		再現率	116	推論規則	165
経験損失最小化	230	最小ルール	199, 202	スケーリング	90
計算幾何学	235	最大重みマッチング問題		スコープ	203
計算言語学	2		119	スタック	151
形態素解析器	77	最大流量問題	120	ステミング	54
桁あふれ	73, 90	最長一致	14	スパースな素性	249
結合	108	最適化	33, 229	スパン	199
結合性	144	最尤推定	31, 64	スパン枝刈り	219
言語モデル	29, 62, 127	雑音のある通信路モデル	27	スピアマンの ρ	55
言語モデル確率	62	座標上昇法	101		
ケンドールの τ	55	サポートベクトルマシン	113	【せ】	
言明	165	参照訳	39, 47	正規文法	178
				生成処理	15
【こ】		【し】		生成モデル	29
語彙化並び替えモデル	134	識別学習	35, 117	正則化	230
		事後確率正規化学習	112	正則化項	117

性能指数	217	タスク指向型評価	39	同期文脈自由文法	175
積集合	178	多値分類問題	260	統計的機械翻訳	4, 27
接合	24, 193	単位元	145	統語論的機械翻訳	207
絶対割引法	68	単一始点最短路問題	147	導出	29, 229
接着ルール	182	単語誤り率	48	動的計画法	
接頭辞	143	単語アライメント	4, 81, 83		33, 48, 88, 138, 164
接尾辞	140	単項	167	凸包	235
セミリング	144	探索	33	トライ	76, 155
線形計画法	119	探索誤り	33	トランスファー方式	10
線形順序付け問題	159	探索空間	33, 139	貪欲法	33, 96, 238
線形補間	66	探索グラフ	142		
線形モデル	34	単調	134		
線形 BLEU	243, 244	単調翻訳	149	【な】	
前件	165	断片化ペナルティ	53	内積	143
潜在変数	30, 229			内的評価	39
線分探索	233	【ち】		長さモデル	85
		置換	24	並び替え	212
【そ】		逐次最適化	240	並び替え制限	149
双対座標下降法	240	知識に基づく機械翻訳	2	並び替え制約	149
ソース頂点	167	中間言語	9		
素性関数	34	忠実さ	40	【に】	
素性ベクトル	34	調整	33, 229	二次計画問題ソルバ	119
外側アルゴリズム	169	頂点	143, 166	二段解析アルゴリズム	223
外側スコア	153	超辺	166	二値分類器	241
外側性能指数	218	直接翻訳	10	二部グラフ	119
ソフトマックス損失	237	チョムスキー標準形		二分化	177, 211
損失関数	230		164, 176	二分化可能同期文脈自由	
損失最小化	230	【て】		文法	211
損失増補推定	254	テイラー展開	244	二分化森	204
		データに基づく機械翻訳	18	ニュートン法	117
【た】		適合率	116	人間翻訳編集率	44
ターゲット頂点	166	デコード	29	【ね】	
ダイクストラ法	147	テストデータ	33	稔性	91
対数線形モデル	34, 117, 130	天井値	94	稔性に基づくモデル	84, 91
対数尤度	73, 117			稔性モデル	92
対数尤度比	81	【と】		【は】	
ダイス係数	81	同期木接合文法	193	パーセプトロンアルゴリ	
大胆な更新	238	同期木置換文法	193	ズム	251
対訳コーパス	17	同期構文解析	214	パープレキシティ	73
対訳辞書	14	同期的更新	259	バックオフ	71
対訳データ	17	同期二分化	211	ハッシュ	76
対訳テキスト	17	同期文法	159, 162	バッチ学習	228
対訳文	17				

幅優先順序	210	部分仮説	138	前向きアルゴリズム	146
汎化誤差	33	ブラックボックス評価	39	前向き後ろ向きアルゴリズム	88, 146
半環	144	ブルーニング	151	マルコフモデル	85
半環構文解析	169	フレイズテーブル	129		
反転トランスダクション文法	215	フレイズペア	126		
		不連続	134	【み】	
【ひ】		フロンティア	193	見出し語化	11
ピアソンの積率相関係数	57	文頭記号	63	未知語	72
ヒープ	148, 151, 172	分配性	144	未定義語	72, 136
ビーム探索	33, 150	文末記号	63	ミニバッチ	228, 248
ビーム幅枝刈り	151	文脈	8		
非局所的な素性	142, 189	文脈自由言語	163	【め】	
非終端記号	162	文脈自由文法	162	メタ評価	57
ヒストグラム枝刈り	151	分野適応	75		
歪み制限	149	分類類似度	20	【も】	
歪み制約	149			目的言語	45
歪みモデル	93	【へ】		モデル誤り	33
ビタビライメント	83	ペアランク最適化	241		
ビタビ近似	32	平滑化	66	【ゆ】	
ビタビ半環	147	平均化パーセプトロン	252	有限状態オートマトン	163, 178
ビタビ翻訳	33, 127	ベイズの定理	28	有限状態トランスデューサ	178
非同期的更新	259	ベイズリスク最小化学習	242		
人手による評価	40	閉包	200	優先度付きキュー	148, 151, 172
被覆ベクトル	139	辺	143	尤度	64, 72
ヒューリスティック関数	153	編集距離	19, 48		
評価者間一致度	46			【よ】	
評価者内一致度	46	【ほ】		要素合成原理	8
評価データ	39	ポアソン分布	85	用例に基づく機械翻訳	4, 17
品詞推定	11	方向	134	用例の大きさ	24
品詞推定器	77	包絡線	235		
ヒンジ損失	241	補完ライメントスパン	199	【ら】	
		補助係数	66	ラグランジュ乗数	104
【ふ】		補助関数	140	ラグランジュの未定乗数法	103
フィッシャーの正確確率検定	130	翻訳の確からしさ	25	ラティス	144
ブートストラップ法	58	翻訳ピラミッド	9	ランキング	42
不可逆圧縮	76	翻訳編集率	44, 50	ランキング問題	241
深さ優先順序	209	翻訳メモリ	19	ランク	176
深さ優先探索	146	翻訳モデル	28, 82		
不完全	94	翻訳森	187	【り】	
復号化	29			流暢さ	40
符号化	29	【ま】			
		マージン最大化	239		
		マージン最大化学習	120		

【る】		レベル順序	209	【わ】	
類似度	25	【ろ】		割引値	69
ルールテーブル	180	ロジスティック損失	241		
【れ】		論理式	165		
レーベンシュタイン距離	48				

【A】		Bayes' theorem	28	cept	94
absolute discounting	68	beam search	33, 150	CFG	162
additive smoothing	71	beam width pruning	151	CG	100, 238
adequacy	40	bilingual corpus	17	Chomsky normal form	164
adjunction	24	bilingual data	17	closure	200
admissible node	201	bilingual dictionary	14	clustering algorithm	77
AER	115	bilingual evaluation		CNF	164, 176
agenda	220	understudy	51	Cocke-Younger-Kasami	
agglomerative clustering	77	bilingual sentence	17	アルゴリズム	164
alignment error rate	115	bilingual text	17	combinatory categorial	
alignment model	85	binarizable-SCFG	211	grammar	182
alignment span	199	binarization	177	communication	38
ALPAC レポート	2	binarized forest	204	commutative	144
analysis	10	binary classifier	241	complement alignment	
and-頂点	168	bipartite graph	119	span	199
and-vertex	168	bitext	17	composed rule	202
and/or グラフ	168	BITG	216	composition	178
and/or graph	168	black box evaluation	39	compositionality principle	
antecedent	165	BLEU	51		8
arity	167	bold update	238	computational geometry	
assimilation	3	bootstrapping	58		235
associative	144	bottom-up	164	computational linguistics	2
asynchronous update	259	bracketing ITG	216	conditional random	
auxiliary function	140	brevity penalty	52	fields	118
averaged perceptron	252	bSCFG	211	confidence interval	58
axiom	165	【C】		conjugate gradient meth-	
A*探索	153	cardinality	140	od	100
A* search	153	case frame	15	consequent	165
【B】		case-based reasoning	18	consistent	201
backward algorithm	146	CBMT	18	context	8
back-off	71	CBR	18	context free grammar	162
batch learning	228	CCG	182	context free language	163
		ceil	94	convex hull	235
				cooling	243

coordinate ascent	101			FST	178
corpus-based machine		[E]		future score	153
translation	18	E ステップ	101	F-measure	53, 115
coverage	74	Earley's algorithm	166		
coverage vector	139	EBMT	4, 17	[G]	
CRF	118	edge	143	Galley-Hopkins-Knight-	
cube pruning	190	edit distance	19	Marcu アルゴリズム	199
CYK アルゴリズム	164	elementary tree	193	generalization error	33
CYK+アルゴリズム	166	EM アルゴリズム	100, 101	generation	15
		encode	29	generative model	29
[D]		entropy	73	GHKM アルゴリズム	199
data-driven machine		entropy pruning	130	glass box evaluation	39
translation	18	envelope	235	glue rule	182
decode	29	Eppstein's algorithm	147	goal	165
deduction system	165	error function	230	goal node	143
deficient	94	evaluation data	39	Good-Turing 法	71
dependency parsing	11	example-based machine		gradient ascent method	
depth-first search	146	translation	4		100
derivation	29	exchange algorithm	78	grading	40
deterministic annealing		expectation maximization		greedy method	33
	243	algorithm	100		
development data	35	expected BLEU	243	[H]	
development set	35	expected loss	242	hash	76
dice coefficient	81	extrinsic evaluation	39	head	143
Dijkstra's algorithm	147			head final language	157
direct translation	10	[F]		head initial language	157
discontinuous	134	F 値	53, 115	head-driven phrase struc-	
discount	69	feature function	34	ture grammar	157
discriminative learning	35	feature vector	34	head-switch	25
dissemination	3	fertility	91	heap	148
distortion constraint	149	fertility model	92	Held-Karp アルゴリズム	
distortion limit	149	figure of merit	217		138
distortion model	93	finite state automaton	163	heuristic function	153
distributional	144	finite state transducer	178	hidden Markov alignment	
distributional similarity	20	Fisher's exact test	130	model	88
domain adaptation	75	fluency	40	hidden variable	30
dot product	143	formula	165	hierarchical phrase	180
Downhill-Simplex 法	232	forward algorithm	146	hierarchical phrase gram-	
dual coordinate descent		forward-backward algo-		mar	180
	240	rithm	88, 146	hierarchical rule	180
dynamic programming	33	fragmentation penalty	53	Hiero 文法	180
		frontier	193	Hiero ルール	180
		FSA	163, 178	Hiero grammar	180

Hiero rule	180	Kendall's τ	55	log likelihood ratio	81
hinge loss	241	KL ダイバージェンス	101	log linear model	34
histogram pruning	151	Kneser–Ney 法	70	longest match	14
HMM	88	knowledge-based machine		LOP	159
HPSG	157	translation	2	lossless compression	76
HTER	44	Kronecker's delta	91	lossy compression	76
human evaluation	40	Kullback–Leibler diver-		loss augmented inference	
human-targeted transla-		gence	101		254
tion edit rate	44			loss function	230
hyperedge	166			L-BFGS 法	117, 238
		[L]		L_1 正則化	230
[I]		Lagrange multiplier	104	L_1/L_2 正則化	260
IBM 制約	150	language model	29	L_2 正則化	230
IBM モデル	81, 82	large margin	120		
IBM constraint	150	latent variable	30		
identity element	145	lattice	144	[M]	
inference rule	165	learning	31	M ステップ	101
inside algorithm	169	learning rate	254	machine translation pyra-	
inside-outside algorithm		lemmatization	11	mid	9
	169	length model	85	MapReduce	258
interlingua	9	Levenshtein distance	48	margin infused relax algo-	
interpolation coefficient	66	lexicalized reordering		rithm	252
intersection	108, 178	model	134	Markov model	85
inter-annotator agreement		lexical translation model		maximum entropy	117
	46	lexicon model	85	maximum flow problem	120
intra-annotator agreement		likelihood	64	maximum likelihood esti-	
	46	limited-memory Broyden–		mation	31
intrinsic evaluation	39	Fletcher–Goldfarb–		maximum weight matching	
inversion transduction		Shanno	117	problem	119
grammar	215	linear BLEU	243	max margin	239
inverted alignment	91	linear interpolation	66	MERT	232
inverted alignment model		linear model	34	meta evaluation	57
	92	linear ordering problem	159	METEOR	53
in-order	210	linear programming	119	method of Lagrange	
item	165	line search	233	multiplier	103
ITG	215	LM	29, 62	minimum Bayes risk	242
		local feature	142	minimum error rate	
[J]		local update	238	training	232
Jensen's inequality	100	logistic loss	241	minimum rule	199
		LogProb 半環	145	mini batch	248
[K]		LogReal 半環	145	MIRA	252
kappa coefficient	46	logsumexp	145	model error	33
KBMT	2	log likelihood	73	modified absolute dis-	
				counting	70

modified Kneser–Ney	71	parse tree	168	
monotone	134	parsing	11	[Q]
monotone translation	149	part-of-speech tagging	11	QP ソルバ
morphological analyzer	77	PBMT	4	119
multi-class classification		Pearson product-moment		quadratic programming
problem	260	correlation coefficient	57	solver
		pegged alignment	97	119
[N]		PER	49	quasi-Newton method
neighbor	96	perceptron algorithm	251	100
neighborhood alignment	96	permutation	212	[R]
Nelder–Mead 法	232	perplexity	73	rank
Newton’s method	117	phrase-based machine		176
noisy channel model	28	translation	4	ranking
noisy-or モデル	133	phrasal ITG	216	42
non-local feature	142, 189	phrase distortion model		rank correlation coefficient
non-terminal	162		127	55
NULL 挿入モデル	92	phrase ITG	216	recall
NULL insertion model	92	phrase pair	126	116
<i>n</i> -gram モデル	62	phrase structure analysis	11	reduce 操作
		phrase structure parsing	11	214
		phrase table	129	reference translation
		phrase translation model		39
			127	regularization
		piecewise linear	234	230
		Poisson distribution	85	regularized empirical risk
		position independent		minimization
		word error rate	49	230
		possible alignment	115	regularizer
		posterior regularization	112	117
		post-editing	3	regular grammar
		post-order	167	178
		POS tagger	77	reordering constraint
		Powell 法	232	149
		precision	116	reordering limit
		predicate-argument struc-		149
		ture analysis	12	RIBES
		prefix	143	55
		pre-reordering method	156	ring
		priority queue	148	145
		PRO	241	risk minimization
		pruning	151	230
		pseudo corpus	249	root
				23
				rule table
				180
				[S]
				S アライメント
				115
				SAMT
				182
				scaling
				90
				SCFG
				175
				scope
				203
				search
				33
				search error
				33
				search graph
				143
				search space
				33
				semantic role labeling
				12
				semiring
				144
				semiring parsing
				169
				sequential minimization
				240
				optimization
				240

[N]

[O]

[P]

[Q]

[R]

[S]

SGD	118, 238, 254	swap	134	tree substitution grammar	193
shard	258	sweep line アルゴリズム	235	tree-based translation	207
shift 操作	213	synchronized update	259	tree-to-string 翻訳モデル	196
shift-reduce アルゴリズム	213	synchronous binarization	211	tree-to-tree 翻訳モデル	196
single source shortest path problem	147	synchronous grammar	159	trie	76
SMO	240	synchronous parsing	214	Tropical 半環	145
smoothing	66	synchronous tree adjoining grammar	193	TSG	193
SMT	4, 27	synchronous tree substitution grammar	193	tuning	33
softmax loss	237	synchronous-CFG	175	two parse algorithm	223
source vertex	167	syntactic transfer	10		
span	199	syntax-augmented machine translation	182	【U】	
span pruning	219	syntax-based translation	207	unary	167
sparse feature	249			underflow	73
Spearman's ρ	55			union	108
stack	151			unsupervised learning	32
STAG	193				
start node	143	【T】		【V】	
statement	165	TAG	24, 193	Vauquois のトライアングル	9
statistical machine translation	4	tail	143	Vauquois triangle	9
stemming	54	target vertex	166	vertex	143
stochastic gradient descent	118	task oriented evaluation	39	Viterbi alignment	83
string-to-string 翻訳モデル	196	Taylor expansion	244	Viterbi approximation	32
string-to-tree 翻訳モデル	196	TER	44, 50	Viterbi semiring	147
structured output learning	35	terminal	162	Viterbi translation	33
structured ramp loss	254	test data	33		
STSG	193	thesaurus	20	【W】	
subjective evaluation	40	TM	28, 82	weight	143
substitution	24	topological order	146	weighted acyclic directed graph	142
suffix	140	top-down	172	weighted acyclic directed hypergraph	166
supervised learning	32	transfer	12	WER	48
support vector machine	113, 240	translation edit rate	44	Witten-Bell 法	67
sure alignment	115	translation forest	187	WordNet	5, 54
SVM	113, 240	translation memory	19	word alignment	81, 83
		translation model	28	word error rate	48
		traveling salesman problem	138	word sense disambiguation	5
		tree adjoining grammar	24		

— 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 ATR 音声言語コミュニケーション研究所
～05年
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 NTT コミュニケーション科学基礎研究所
～08年
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 NTT ソフトウェア株式会社勤務
～98年
2000年 株式会社国際電気通信基礎技術研究所
～06年
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 兵庫県庁勤務
～08年
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

© Watanabe, Imamura, Kazawa, Neubig, Nakazawa 2014

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



本書のコピー、スキャン、デジタル化等の無断複製・転載は著作権法上での例外を除き禁じられています。購入者以外の第三者による本書の電子データ化及び電子書籍化は、いかなる場合も認めておりません。

落丁・乱丁本はお取替えいたします