革新的発展期を迎えた統計翻訳

永田 昌明（ながた まさあき）/須藤 克仁（すどう かつひと）
鈴木 潤（すずき じゅん）/秋葉 泰弘（あきば やすひろ）
平尾 努（ひらお つとむ）/塚田 元（つかだ はじめ）
NTTコミュニケーション科学基礎研究所

英語と日本語は語順の対応関係が正順になる部分と逆順になる部分が複雑に入り混じっており，おそらく世界でもっとも翻訳が難しい言語対の１つです．私たちは，主辞後置性という日本語の言語学的な特徴を利用して，英語の文の単語を日本語と同じ語順に並べ替えてから日本語へ翻訳する方法を考案し，英日翻訳の精度を劇的に改善しました．さらにこの方法は，中国語から日本語への翻訳においても非常に有効であることも確認しました．

機械翻訳はかなわぬ夢？

コンピュータを利用して，ある言語を別の言語に翻訳する技術を「機械翻訳」と呼びます．機械翻訳の研究はコンピュータの誕生とほぼ同時の1950年代から始まり，今日までに多数の機械翻訳システムが開発されました．

近年インターネットの普及により，ごく普通のユーザが英語・中国語・韓国語などの外国語で書かれたWebページに接する機会が飛躍的に増えました．世界中に市場を持つ多国籍企業は，マニュアルなどの製品情報を迅速，かつ正確に現地語で提供する必要があります．言葉の壁を越えて人とコミュニケーションしたい，あるいは，言葉の壁を越えて知識を交換したいという想いは人類の普遍的な要求の１つといってよいでしょう．しかし，正直なところ従来の機械翻訳システムがこれらのさまざまなユーザの要求を満足させてきたとはいい難いのではないかと思います．

ルールベース翻訳から統計翻訳へ

従来の機械翻訳システムでは，新しい言語間の翻訳を実現するために，数人から数十人の専門家が何年もの歳月をかけて翻訳規則や対訳辞書を人手で作成していました．このような機械翻訳へのアプローチを「ルールベース翻訳」と呼びます．ルールベース翻訳は人手作業による精度の限界に到達し，さらなる改善は難しいといわれています．これに対して「統計翻訳」は，数10万～数100万文の大規模な対訳データから，翻訳規則や対訳辞書に相当する統計モデルを自動的に学習し，新しい言語対や特定の分野の機械翻訳システムを短期間に低コストで作成することを目指す，発展途上の技術です．統計翻訳の概要を**図１**に示します．

統計翻訳は，英語とフランス語の両

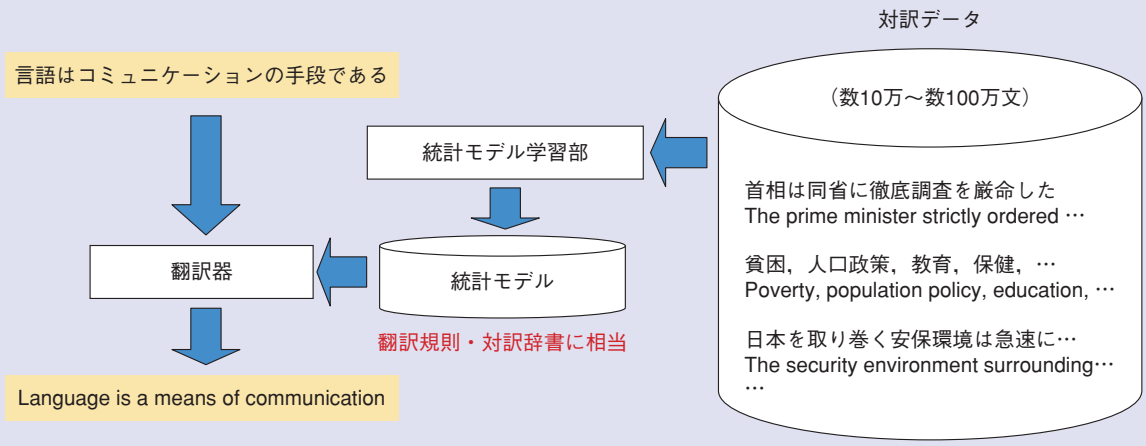


図１　統計翻訳の概要

方で出版されているカナダの国会議事録「ハンサード」を使って，1990年前後にIBMが英語とフランス語の間の機械翻訳システムを作成したのが最初の試みです．2000年代には翻訳の基本単位を単語から句に拡張する「句に基づく翻訳」の研究が進み，英語とフランス語のような語順が近い言語対では実用レベルに到達しました．

2005年前後からは自然言語の階層構造や構文理論を利用する「木に基づく翻訳」の研究が進み，英語とフランス語の翻訳に比べて語順が大きく異なることが問題である英語と中国語の翻訳でも，統計翻訳の精度がルールベース翻訳より高いことが確実になりました．しかし，英語と中国語に比べてさらに大きく語順が異なる英語と日本語の翻訳では，統計翻訳の精度は従来のルールベース翻訳の精度を上回ることができませんでした．

事前並べ替え翻訳

原言語（翻訳元）の文の単語を目的言語（翻訳先）の語順に並べ替えてから統計翻訳を行う「事前並べ替え」というアイデアは，2000年代前半から存在していましたが，2010年前後からGoogle，Microsoft，IBM，NTTなど世界の主要な研究機関は，語順の違いを克服する技術としての事前並べ替えに注目し始めました．事前並べ替えは，原言語の文を構文解析して得られた構文構造に対し，「並べ替え規則」を適用して目的言語の語順に変換します．並べ替え規則は人手で作成する場合が多いですが，単語対応付きの対訳データから自動学習する方法も提案されています．

このような状況の中で，NTTは主辞後置性と呼ばれる日本語の言語学的な特徴に着目し，「主辞を後置する」（主辞後置化）というただ1つの規則を使って英語の単語を日本語の語順に並べ替える方法を考案し，英日翻訳の精度を劇的に改善しました[(1)]．2011年に開催された評価型ワークショップNTCIR-9の特許翻訳タスクの英日翻訳において，NTTと東京大学の共同チームは高精度な英語構文解析器Enjuと，主辞後置化に基づく事前並べ替えを組み合わせることにより，トップの成績を収めました．人手による英日翻訳の精度評価で統計翻訳がルールベース翻訳を上回ったのは，これが史上初めてです[(2)，(3)]．

日本語の主辞後置性に基づく事前並べ替え

日本語の主辞後置性に基づく事前並べ替えの概要を**図2**に示します．主辞とは，文を構成する部品である句において，句の文法的な役割を決める単語です．例えば前置詞句なら前置詞が主辞です．別の言い方をすると，小学校の国語の授業に習う「係り受け」において係り先になる単語が主辞です．日本語は「必ず前から後ろへ係る」，すなわち修飾先の単語が必ず文の後ろにあります．これが「主辞後置性」です．

実は，日本語ほど厳密な主辞後置性を持つ言語は世界の中でも稀です．一般的には，図2の英語や中国語の例のように「前から後ろ」にも「後ろから前」にも修飾します．例えば，英語は動詞に対して主語は前から，目的語は後ろから修飾します．また名詞に対して形容詞は前から，前置詞は後ろから修飾します．中国語も動詞に対して主語は前から，目的語は後ろから修飾しますが，名詞に対しては基本的に前から修飾します．

この日本語の主辞後置性のおかげで，翻訳元の言語（英語や中国語）の構文構造（係り受け関係）に基づいて，「必ず前から後ろへ係る」ように単語の順番を入れ替えると，翻訳元の言語を日本語と同じ語順に変換できます．これが主辞後置化による事前並べ替えです．語順が同じになれば，あとは逐語訳するだけなので，非常に高精度な翻訳が実現可能になります．日本語の主辞後置性に基づく事前並べ替えは，翻訳先の言語である日本語の性質だけを用いるので，翻訳元の言語の構文構造が分かっていれば，どんな

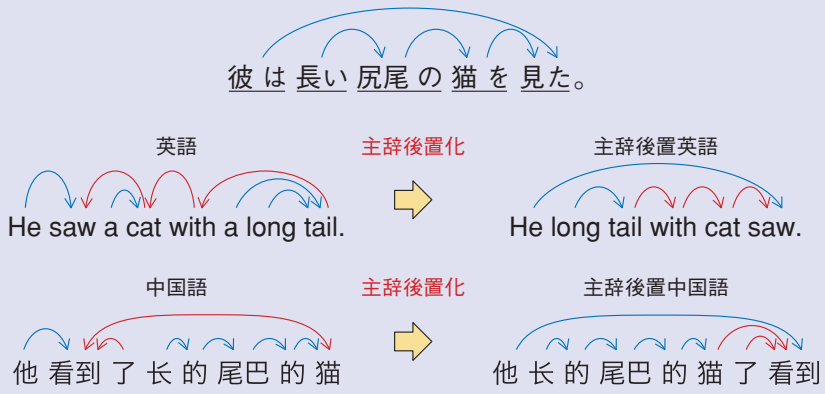


図2　日本語の主辞後置性に基づく事前並べ替え

言語へも適用可能です．

一方，日本語を外国語（英語や中国語）へ翻訳する場合，日本語の構文構造において「後ろから前へ」反転させる係り受け関係を翻訳先の言語に応じて選択する必要があります．この問題を解決しなければならないため，日本語から外国語への翻訳は，外国語から日本語への翻訳に比べて難しいのです．

技術文書の多言語翻訳

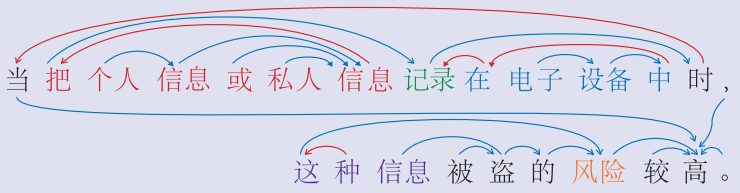
私たちは，事前並べ替え方式による外国語から日本語への統計翻訳の実現可能性を検証するために，特許文書を対象として，英語・中国語・韓国語から日本語への統計翻訳システムを作成しました．特許には「パテントファミリー」と呼ばれるものがあります．これは同じ発明を複数の国へ出願するために，1つの特許出願に対して優先権を主張して各国へ出願した「特許出願のまとまり」のことです．パテントファミリーは完全な対訳ではありませんが，対訳になっている部分を多く含んでいるので，パテントファミリーをマイニングすることにより大規模な対訳データを抽出することができます．私たちは，2004年以降の日本，米国，中国，韓国の特許文書から，英語—日本語（約300万文），中国語—日本語（約800万文），韓国語—日本語（約200万文）の対訳データを作成しました．特に中日と韓日の特許対訳データは，私たちの知る限り，この言語対に関する世界最大の対訳データです．

NTTが考案した日本語の主辞後置性に基づく事前並べ替えを適用するためには，翻訳元の言語の文の構文構造（係り受け関係）を高精度で解析する技術が必要です．私たちは，英語（新聞記事4万文，特許1万文）と中国語（新聞記事5万文，特許2万文）の学習データを用意し，「半教師あり学習」による係り受け解析モデルの学習技術を利用して，英語と中国語の係り受け解析器を作成しました．この半教師あり学習による係り受け解析モデルの学習技術は，2009年にNTTが考案したもので，係り受け解析に関する国際標準ベンチマークテストデータ（英語，チェコ語）でトップの成績を収めています[4]．

図3に中国語の特許文の係り受け解析の例，**図4**に特許文の中日翻訳の例を示します．一般に特許文は非常に長く，複雑な係り受け構造を持っていますが，中国語の係り受け構造を正しく解析できれば，主辞後置化によって中国語における修飾・被修飾の関係を正確に反映した日本語を生成できることが分かります．なお韓国語は日本語とほぼ語順が同じなので，韓日翻訳には事前並べ替えを適用していません．

翻訳精度の自動評価

最後に，翻訳精度の自動評価について簡単に説明します．機械翻訳の精度を客観的に計測することは実は非常に難しい問題です．ある文に対する翻訳の正解は何通りもあり，訳語選択



個人 情報 や 個人 情報 を 電子 機器 に 記録 する 際 に は 、
この よう な 情報 が 盗難 さ れる リスク が 高い 。

図3　中国語の特許文の係り受け解析の例

原文

当把个人信息或私人信息记录在电子设备中时，
这种信息被盗的风险较高。

並べ替え結果

个人 信息 或 私人 信息 把 电子 设备 中 在 记录 时 当 ，
种 这 信息 が 被 盗 的 风险 较 高 。

翻訳結果（NTT方式）

個人 情報 や 個人 情報 を 電子 機器 に 記録 する 際 に は 、
この よう な 情報 が 盗難 さ れる リスク が 高い 。

翻訳結果（事前並び替えなし）

また 、 個人 情報 や 個人 情報 が 記録 さ れる 際 に 、
電子 機器 に この よう な 情報 が 盗聴 さ れる 危険 性 が 高い 。

参照訳

電子機器に個人情報やプライバシーに関わる情報が記録されている場合には、その様な情報を盗み取られるリスクが高い

図4　特許文の中日翻訳の例

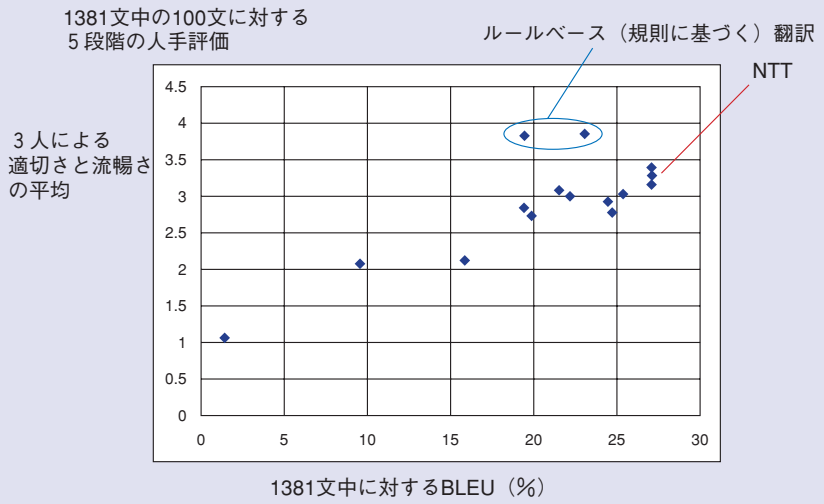


図5　NTCIR-7 日英特許翻訳タスク（2008）におけるBLEUと人手評価の関係

の誤りと語順の誤りのどちらを重視するかは主観的な判断です．1990年代に考案された機械翻訳の自動評価尺度BLEU（BiLingual Evaluation Understudy）は，機械翻訳の研究の活性化に大きく貢献しました．これは食味計の発明によりお米のおいしさを数値化することで，産地間の競争や品種改良が進むのに似ています．しかし，BLEUは日本語と英語の翻訳では人手による評価とあまり一致しないという問題がありました．2008年に開催された評価型ワークショップNTCIR-7における自動評価尺度BLEUと人手による評価の相関関係を**図5**に示します．

そこで私たちは，翻訳結果と正解の語順の一致度を重視する自動評価尺度RIBES（Rank-based Intuitive Bilingual Evaluation Score：ライビーズ）を考案し，オープンソースソフトウェアとして公開しています(5)．

RIBESは，前述のNTCIR-9において自動評価尺度の1つとして採用され，主催者の評価においても，英日・日英・中英翻訳タスクにおいて人間による評価との相関がBLEUより高いことが示されました(6)．

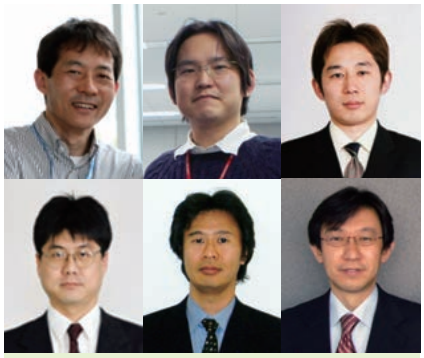
今こそ機械翻訳の実用化へ

特許，マニュアル，科学技術論文などの技術文書では，客観的・論理的な内容の伝達，すなわち翻訳元の言語における修飾・被修飾の関係を翻訳先の言語に反映することが，意味を正確に伝達するうえでもっとも重要です．私たちは，特許に代表されるような，100万文を超える対訳データが得られる分野において，外国語から日本語への統計翻訳は実用レベルに到達したと考えています．

一方，日本語から英語への翻訳については，その差は縮まっていますが，統計翻訳の精度は依然として従来のルールベースの翻訳の精度を上回っていません．今後は，日本語から外国語への翻訳の精度向上に取り組むとともに，技術文書からビジネス文書，話し言葉へと翻訳対象を広げていきたいと考えています．

■参考文献

(1) H. Isozaki, K. Sudoh, H. Tsukada, and K. Duh：“HPSG-Based Preprocessing for English-to-Japanese Translation,” ACM TALIP, Vol.11, No.3, 2012.
(2) I. Goto, B. Lu, K. P. Chow, E. Sumita, and B. K. Tsou：“Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop,” NTCIR-9, pp.559-578, Tokyo, Japan, Dec. 2011.
(3) K. Sudoh, K. Duh, H. Tsukada, M. Nagata, X. Wu, T. Matsuzaki, and J. Tsujii：“NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT,” NTCIR-9, pp.585-592, Tokyo, Japan, Dec. 2011.
(4) J. Suzuki, H. Isozaki, X. Carreras, and M. Collins：“An Empirical Study of Semi-supervised Structured Conditional Models for dependency Parsing,” EMNLP 2009, pp.551-560, Suntec, Singapore, Aug. 2009.
(5) http://www.kecl.ntt.co.jp/icl/lirg/ribes/index-j.html
(6) H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada：“Automatic Evaluation of Translation Quality for Distant Language Pairs,” EMNLP 2010, pp.944-952, Cambridge, U.S.A., Oct. 2010.



（上段左から）永田　昌明/ 須藤　克仁/ 鈴木　潤
（下段左から）秋葉　泰弘/ 平尾　努/ 塚田　元

外国語から日本語への翻訳において優れた独自技術を開発できたという高揚感が本稿から伝われば幸いです．一方，日本語から外国語への翻訳は統計翻訳において残された最大の難問であり，今後はこれに真摯に取り組みたいと思います．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
協創情報研究部　言語知能研究グループ
TEL　0774-93-5149
FAX　0774-93-5345
E-mail　nagata.masaaki@lab.ntt.co.jp