別添１

機械翻訳の現状調査結果

機械翻訳の現状調査結果　目次

機械翻訳の現状調査

機械翻訳の現状調査は機械翻訳の翻訳方式、翻訳前と翻訳後の処理、翻訳対象の言語、翻訳の精度などの機械翻訳の機能・特性の現状について、文献、書籍、インターネット情報、各翻訳方式の研究者の知見を調査して、特許庁の将来の機械翻訳システムの導入にあたり参考となる情報を収集、整理、分析した。

翻訳対象の言語対は日本語⇔英語、中国語⇔日本語、韓国語⇔日本語、ベトナム語⇔日本語、タイ語⇔日本語及びインドネシア語⇔日本語の６つである。

調査対象は、①翻訳方式、②言語リソース、③要素技術、④翻訳精度、⑤アプリケーションとの連携、そして⑥言語別の状況とした。

① 「翻訳方式」は、統計的機械翻訳、ルールベース機械翻訳、用例ベース翻訳、ハイブリッド翻訳そしてピボット翻訳の５つの翻訳方式について、各方式のメリット・デメリット、翻訳精度の現状、新たな言語対の追加容易性について説明する。
なお、調査期間中にGoogle等からリリースされたニューラルネットワークを用いた機械翻訳システムについては、「機械翻訳の利用及び将来性に係る調査」を参照のこと。

② 「言語リソース」は、機械翻訳システムの開発に不可欠な各言語のリソースの調査である。翻訳対象の言語（日本語、英語、中国語、韓国語、ベトナム語、タイ語、インドネシア語）について、利用できる辞書、単言語コーパス、対訳コーパスについて説明する。

③ 「要素技術」は、機械翻訳システムの実現に不可欠な各言語を扱うツールの調査である。翻訳対象の言語（日本語、英語、中国語、韓国語、ベトナム語、タイ語、インドネシア語）について、形態素解析、構文解析、意味解析、文アライメント作成、そして統計翻訳で必要な言語モデルと学習・デコーダの有無と状況について説明する。

④ 「精度評価」は、機械翻訳の精度評価に関する調査である。翻訳精度の評価手法である人手評価及び自動評価の手法について、評価方法とその特徴を説明する。また、過去の機械翻訳結果の評価について説明する。

⑤ 「アプリケーションとの連携」は、XML形式の特許文書を翻訳する際に考慮すべき事項等、機械翻訳システムが他のアプリケーションと連携する際の課題について説

明する。

⑥　「言語別の状況」は、各言語の特徴と、前記②言語リソース、③要素技術との関係等について調査した結果を説明する。

1. 翻訳方式

1.1 統計翻訳（SMT）

1.1.1 SMT の概要

(1) 機械翻訳手法のパラダイムシフト

統計翻訳（Statistical Machine Translation,SMT）とは、原文と訳文を大量に集めた対訳コーパスと統計的な学習アルゴリズムにより、翻訳システムを自動的に構築する手法である。提案されたのは 1988 年であるが[1]、以下の理由で、当時は大きく発展することはなかった。

① 対訳コーパスの量や計算機の能力が足りなかったため、高い翻訳精度が出なかったこと。
② Word-Based と呼ばれた当時の手法は、英語とフランス語のように語順が似ており、語彙が相互に関連する言語対には適用できたが、英語と日本語のように文法が著しく異なり、語彙にほとんど重なりがない言語対には適用できなかったこと。

当時は、ルールベース翻訳（Rule-Based Machine Translation,RBMT）の方が明らかに高精度で、汎用性があった。

2000 年前後に、SMT の改良版であるフレーズベース統計翻訳（Phrase-Based Statistical Machine Translation,PBSMT）が提案された。翻訳の単位を単語（Word）から句（Phrase）[2]という大きな単位に置き換えて計算量を抑えることにより、翻訳精度を大幅に改善することに成功し、英語と日本語のような翻訳が難しい言語対にも適用可能となった[3]。

対訳コーパスの蓄積、計算機の能力やメモリ容量の増大など研究環境も改善され、SMT が様々な言語対に適用されるなど、研究者の間で急速に広がった。さらに、PBSMT に続き、SMT の新たな方式として hierarchical,tree-to-string,operation-sequence などの各種方式[4]が提案された。

2010 年になると、英語から日本語への翻訳のように文法が著しく異なる言語間の翻訳においても高い翻訳精度を達成できる PRE-ORDERING 方式の統計翻訳が提案され、日本においても SMT の実用化が加速された[5]。これは、日本語の言語特性を利用する方式であり、多

---

1 Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin: A STATISTICAL APPROACH TO MACHINE TRANSLATION. Computational Linguistics, Vol.16, No. 2, pp.79-85, 1990.

2 名詞句や動詞句のようなものであると捉えると分かりやすい。しかしながら、実装はこれらの文法概念と無関係な単語列であった。

3 英日の機械翻訳は、英仏の機械翻訳に比べると、精度は劣る傾向にあった。

4 これらの方式は大きな性能差はないので、本稿では、取り扱わない。

5 Hideki Isozaki, Katsuhito Sudoh, Hajime Tsukada, and Kevin Duh: Head Finalization: A Simple Reordering Rule for SOV Languages, Proceedings of WMT-2010（ACL 2010 Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR）, pp.244--251, 2010.

言語が取り扱えるというSMTの汎用性を犠牲にしたものであるが、従来のSMTと比べ、翻訳精度が飛躍的に向上した。

(2) 単語に基づく統計翻訳（Word-Based SMT）

統計翻訳は、過去の事例をもとに、翻訳文として最も確からしいものを選ぶことである。この統計翻訳の中で最も基本的なものが、単語に基づく統計翻訳である。

例えば[6]、フランス語「il croit」を英語に翻訳する場合を見てみる。「il」には「he」と「it」の二つの訳語が、「croit」には「thinks」と「grows」の二つの訳語があるので、「il croit」の訳文として「he thinks」「he grows」「it thinks」「it grows」の４通りがありうる。一般的に使用されている英文のうち、上記４通りの英文の使用頻度を比べると、「he thinks」が一番高いと考えられる。したがって、統計翻訳においては、「il croit」は「he thinks」に翻訳されることになる。

原文から外国語の単語の集合へ変換する部分は翻訳モデルと呼ばれ、対訳コーパスから統計的に学習される。また、尤もらしい訳文を選ぶ部分は言語モデルと呼ばれ、翻訳言語のデータから学習される。翻訳時には、原文に翻訳モデルと言語モデルを適用して生成した複数の翻訳候補から、最も尤もらしい訳文を選定する。

この方法は、1988年ころの計算機では処理に計算時間がかかりすぎただけでなく、いろいろ制約を付加して計算量を端折ると、結果として十分な精度が出にくく、特に日本語と英語のように文法や語彙・語順が著しく違う言語対の間では、高品質の翻訳が実現出来なかったので普及しなかった。

(3) フレーズに基づく統計翻訳（Phrase-Based SMT）

現在、SMTの標準的な手法と考えられているのが、フレーズに基づく統計翻訳（Phrase-Based SMT,PBSMT）である。まず、翻訳対象の原言語と目的言語の対訳コーパスを用意し、対訳コーパスの中でフレーズの対応関係を自動的に決定し（□で囲まれたフレーズが相互に対訳関係になっている)、対訳コーパスから翻訳確率付きのフレーズの対訳辞書を作成（学習）する。図１．１．１-１に、原言語が英語、目的言語が日本語であり、“I will go”で始まる英文及びそれに対応する日本語文が対となっている対訳コーパス（文対数は５）の例を示す。この場合、図中の□で囲まれたフレーズが対応関係にある。次に、対訳コーパスにおいて、英語の“I will go”が日本語でどのように翻訳されるかを調べる。例では、“I will go”を含む文対５例中、「します」と翻訳されるものが２例、「に行きます」に翻訳されるものが３例である。これらの結果をもとに“I will go”を「します」及び「に行きます」と翻訳する確率を算出すると、それぞれ２／５（＝40%）及び３／５（＝60%）となるから、確率付き対訳辞書は、“I will go”に対する翻訳フレーズである「します」及び「に行きます」に対して、それぞれ翻訳確率を対応づける。

6 Kishore Papineni（Yahoo）による。

このように作成した翻訳確率付きのフレーズの対訳辞書を、フレーズテーブル又は Phrase-Based SMT の翻訳モデルという。

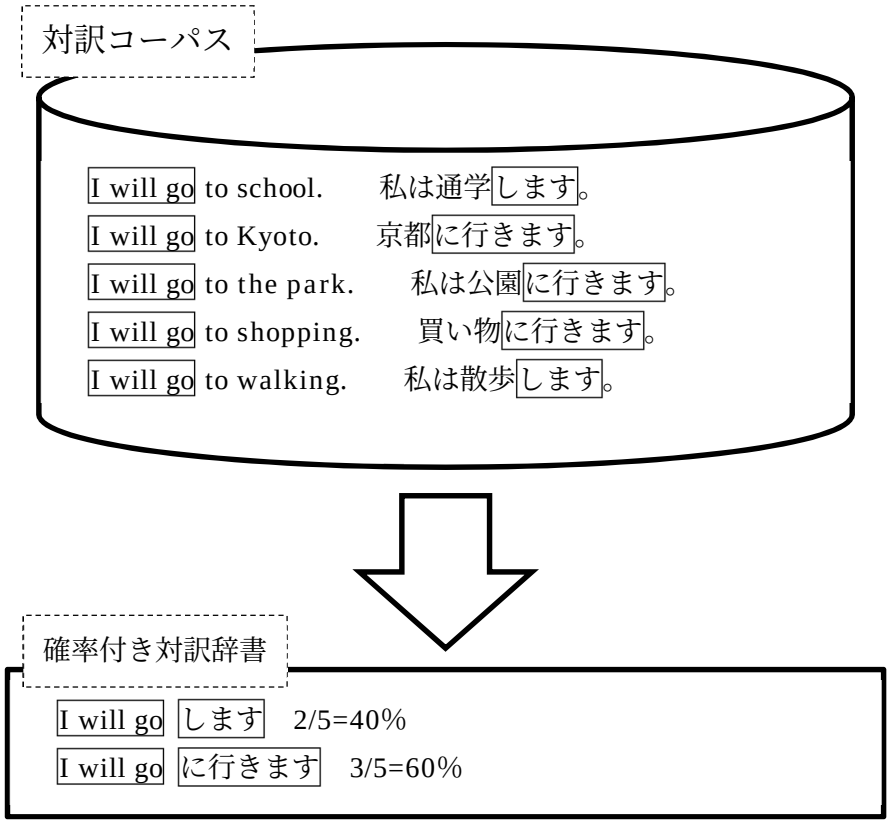


図１．１．１-１　翻訳モデル（フレーズの翻訳確率付き対訳辞書）の自動学習

翻訳時には、原文とフレーズテーブルを対応づけて求めた翻訳確率を参照して、最も尤もらしくなるように翻訳文を生成・選択する。

翻訳のプロセスは以下のとおりになる。

① 入力文をフレーズ[7]に分割する。
② 確率付き対訳辞書における翻訳確率をもとに、フレーズが最も尤もらしくなるように翻訳フレーズを選択する。
③ フレーズの順序を確率的に調整する。

図１．１．１-２に示したように、四角で囲まれたフレーズ毎に翻訳され（例えば、“I will go”が「行きます」に翻訳され)、その後、翻訳されたフレーズの語順が調整される[8]。

[7] 文法的な意味の句でなく、翻訳する上で固定的に取り扱える単語列。
[8] フレーズの内部は固定されており、フレーズの中では語順は変わらない。

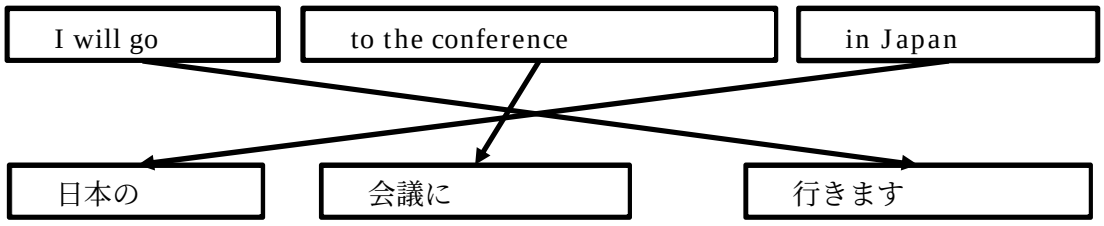


図１．１．１-２　フレーズベース統計翻訳の動作例

(4) PRE-ORDERING 技術

PBSMT では、翻訳言語によっては、フレーズの並べ替えを制御する必要が生じる。例えば、日本語と英語のように文法が異なる言語間の翻訳は、以下に説明する PRE-ORDERING 技術に基づく語順変換処理を行ってフレーズの並び替え処理を行うことにより、翻訳精度を改善できる [5]。

PRE-ORDERING 技術を用いた英日翻訳の処理手順は、以下のようになる。

① 英語文の構文構造を解析する。
② 英語の語順を日本語の語順に変換する。その際に、日本語の基本文形は、「主語 S、目的語 O、動詞 V の順であり、動詞が（より一般的には、主辞）が必ず最後になる」性質を利用する。
③ その後、訳語を選択する。

図１．１．１-３に英語を日本語に翻訳する例を示す。もとの英文を、英語のまま日本語の語順に変換した上で、英語を日本語に変換する。

入力（英語文）：FIG. 3C is a graph illustrating a simulation that includes the effects of resonance, cyclic clocks, and a change in logic current.
語彙英語・語順日本語：FIG. 3C _va1_ resonance of effects , cyclic clocks , and logic current in change _va2_ includes that simulation _va2_ illustrating graph is .
出力（日本語）：図３Ｃは、共振による効果、環状のクロック、および論理電流の変化を含むシミュレーションを示すグラフである。

図１．１．１-３　PRE-ORDERING 技術による翻訳過程

なお、方式が提案された当初は、語順変換は人手で構築された規則が用いられていた

が、その後、語順変換を自動的に学習する手法が提案されている[9]。これにより、特許文書のように文が長く複雑な構文の文が多い文書においても、高精度な多言語翻訳が可能になってきている。

1.1.2 SMT の一般的なメリットとデメリット

SMT のメリットとデメリットを、RBMT などの他の翻訳方式と比較して説明する。

(1) SMT のメリット

① システム構築が容易であること

RBMT と比べた場合の SMT の最大のメリットは、ある程度の量の対訳コーパスを用意すれば翻訳システムが構築できるため、多言語化が容易なことである。RBMT のシステムを構築する場合には、言語処理の専門家が必要であり、大規模なシステムを短期間で開発することは非常に難しい。これに対し SMT の場合には、対訳コーパスを用意すればよいため、翻訳システムの構築作業は単純である。しかも特許分野ではパテントファミリーとして内容の対応が取れている文書が大量にあることから、それらの中から文同士の対応を推定して対訳コーパスを短時間で安価に構築することが可能である。また、翻訳精度を向上させるためには、基本的には対訳コーパスの規模を大きくしていけばよい。後述するように、RBMT の場合には、追加する情報と既にシステムに登録されている情報の間の整合性を考慮する必要があるが、SMT は統計に基づく翻訳モデルと言語モデルを作成するので、そういった問題に自動的に対応できる。

② 多言語化が容易であること

システムの多言語化が比較的容易であるのも SMT の特徴である。RBMT の場合は、翻訳対象とする言語（英日翻訳であれば英語と日本語）ごとに、対訳辞書や構造変換規則などの情報が必要であり、N 個の言語間で相互に翻訳ができるシステムを構築しようとすると、そのような知識を N×（N-1）組用意する必要がある。それに対し SMT を構築する場合、理想的には全ての文に対し N 個の言語での表現を用意すればよい。すなわち、最初に日本語の文セットを用意して、それらを英語、中国語、韓国語に翻訳すれば、４言語の対訳コーパスができるので、これを用いて 12 通りの組み合わせで対訳コーパスを作成して４言語間で相互に翻訳ができる 12 通りの翻訳システムが実現できることになる。

[9] 渡辺太郎,今村賢治,賀沢秀人,Graham Neubig,and 中澤敏明. 機械翻訳. コロナ社, 2014. 5.5.2 事前並び替えモデル

③　訳文が自然であること

RBMT の翻訳結果は、内容が正確であっても、硬い、ぎこちない、機械的で読みにくいといった印象を与える場合が多かったが、SMT では目標言語（英日翻訳であれば日本語）の複数の訳文候補の中から言語モデルを用いてその言語として最も尤もらしい表現が選択されるため、RBMT に比べて自然で読みやすい文が出力されることが多い。この性質は間違いなく SMT の大きなメリットであるが、一方で、翻訳結果が自然過ぎて、誤訳があっても気付きにくいというデメリットもあることには注意する必要がある。

(2) SMT のデメリット

①　必要とする計算資源が大きいこと

SMT は RBMT と比較して、翻訳を実行するのに大量のメモリとより多くの計算処理を必要とする。これはフレーズベースの翻訳モデルと言語モデルが非常に大きくなるためである（二つを合わせて 122GB 程度になる事例もある）。論文などの形で比較結果を公表しているデータはないが、調査者による知見として、日英翻訳で最高精度を目指した場合、SMT は RBMT と比べて、50 倍から 100 倍程度の計算資源を要することが分かっている。翻訳速度は、処理の並列化により改善できる部分もあるため、単純に翻訳時間が 100 倍になるというわけではないが、並列化するための計算資源は必要となる。計算機の価格性能比は常に改善しており、SMT による翻訳システムも既に実用化されてはいるが、大規模な翻訳処理を行う際には、コストの面で SMT を採用できない場合や、計算資源をある範囲内に抑えるために目標とする翻訳精度を少し落として利用したり、対訳コーパスの量を制限したりすることが必要になる場合もありうるという点には注意が必要である。

②　翻訳時の抜けや語の湧き出しがあること

対訳コーパスとして用意される対訳文は、文としての対応は間違っていない場合でも、文に含まれる単語同士の対応が取れない場合も存在する。例えば、日本語では文脈から容易に読み取れる語は省略されるのが一般的である。また、翻訳文を分かりやすくするため、翻訳者が翻訳の過程で原文にはない語を補う場合もある。その結果、翻訳の過程でこうした語の削除や追加が行われた原文と訳文を対訳コーパスとして使用すると、その学習結果により得られた翻訳モデルには誤った対訳情報が入ってしまう。その翻訳モデルを用いた統計翻訳では、原文に存在している語が訳出されなかったり、原文には存在しない語が訳出されたりするという現象が発生することがある。RBMT においても規則や翻訳アルゴリズムの不備により、このような現象が発生することもあるが、極めて稀であり、SMT のように顕著な問題として認識されることはない。

③　文章全体での訳語の統一性がないこと

RBMT が実用化された際に、メリットの一つとして挙げられたのが、訳語が統一される点であった。RBMT の導入前は、一つの文書を複数の翻訳者が分担して翻訳する場合、同じ単語であっても翻訳者によって異なる訳語を使用する可能性があるため、訳語が統一されているかをチェックする必要があった。その点、RBMT では翻訳辞書により特定の語には同一の訳語が用いられるため、仮に間違って翻訳されている場合でも辞書に正しい訳語を登録することで修正が容易であり、訳語を統一することができた。それに対し現在の SMT では、それぞれの単語が、その単語が出現した文脈（フレーズ）で最適な訳が選定される。パテントファミリーの特許文書を用いて対訳コーパスを自動構築した場合、対訳コーパスには出願者によって異なる訳語が用いられている対訳が含まれる可能性がある。その結果、同一の語に対し、文によって異なる訳語が使われる場合が生じ、結果として文書を通読した場合に訳語の統一性がなく、読みにくくなる場合がある。

1.1.3　特許分野での適用上の課題

特許文書は一般に、一文が長く、専門用語が多く含まれ、さらに新語が頻出する等の理由から、機械翻訳が困難と考えられてきた。これらの課題は、翻訳方式を問わず特許の機械翻訳においては重要な課題であるが、専門用語や新語の問題については、パテントファミリーの文書を用いて対訳コーパスを随時拡充していくことができるという前提に立てば、従来の RBMT で必要であった辞書整備作業に比べると、システムの継続的な改善が可能であると考えることができる。

ただし、PRE-ORDERING 技術により原文の語順を変更する方式を採用しているシステムでは、単に対訳文を追加するだけでは期待する形に翻訳品質が改善されない可能性もある。すなわち、単語の区切り位置や係り受けが間違っていると、動詞の把握が難しくなり、語順が適切に変更できないからである。そのような場合には、対訳コーパスとは別に、原言語のタグ付きコーパスを作成して形態素解析や構文解析の精度を改善する必要が生ずる。

1.1.4　多言語化の現状

1.1.2 で述べたように、RBMT に比べて多言語化が容易であることは、SMT を採用するメリットの一つである。旅行会話を対象とした音声翻訳システムにおいては、29 言語間の翻訳システムも実現され、翻訳サービスが提供されている[10]。

多言語翻訳システムの開発にあたっては、各言語間の対訳コーパスを構築する必要があるが、旅行会話用音声翻訳システムのようなケースでは、翻訳される文に含まれる語彙や表現は日常会話で使われる基本的なレベルのものである。そのため、中心となる言語（例えば日本語もしくは英語）でテキストを準備し、それを一般的なスキルを有する翻訳者によって残り全ての言語に翻訳することで、対象言語すべてで用いることができる対訳コー

[10] 多言語音声翻訳アプリ<ボイストラ>, http://voicetra.nict.go.jp/（最終検索日:2016 年 7 月 13 日）

パスを構築することができる。

しかしながら、特許文書は記載内容の専門性が高いため、旅行会話用のものと同じ方法で対訳コーパスを構築するのは現実的ではない。そのため、現在は翻訳対象となる言語対ごとにパテントファミリーの特許文書を用いて対訳コーパスを構築するが、特許の出願件数の違いにより言語対ごとに対訳コーパスのサイズは異なる。なお、アセアン言語は、アセアン諸国特許庁での電子出願の導入と普及が遅れているため、まだ十分なサイズの対訳コーパスは構築できていない。そのため、アセアン言語を対象とした実用レベルの対訳コーパスはまだ存在していないため、アセアン言語の SMT は実現されていない。

1.1.5 翻訳精度の現状、精度向上に向けた取り組み

1.4 節で述べるように、様々な調査研究や評価型ワークショップにおいて、SMT の翻訳精度は一部の言語対を除いて RBMT の精度を上回る状況となってきている。

現時点における日英機械翻訳の傾向を見ると、明細書の短文の翻訳精度については、SMT が RBMT を上回ることが報告されているものの、長文の翻訳精度については RBMT と比べやや劣る傾向にある。また、請求項、拒絶理由通知書、審決といった、対訳コーパスの量が十分にない文書、一般的な文章の形式に近いものの引用など特殊な表現を多く含む文章の機械翻訳に関しては、SMT には課題がある。こういった課題を解決し、翻訳精度を向上させる方法として、例えば、請求項に関しては、請求項に内在するパターンの解析と変換で SMT 翻訳しやすいように翻訳前処理をしたりすることで翻訳精度の向上を図ることが検討されている[11]。

1.1.6 カスタマイズの容易性

機械翻訳におけるカスタマイズとは、一般に、システムが出力する訳文を利用者が期待する形で出力されるようにシステムを調整する作業を指す。例えば、未知語や誤訳を利用者が望む訳語で訳出されるように調整することである。

RBMT におけるカスタマイズとしては、ユーザ辞書への辞書登録を行ったり、利用者や組織単位で構築した複数個のユーザ辞書を、優先順序を指定して用いたりすることが挙げられる。また、市販されている RBMT システムでは、目標言語が日本語の場合には、文体を「ですます調」とするか「である調」とするかといった指定や、句読点を「。、」とするか「.,」とするかといった指定ができることが多い。

SMT においても、対訳コーパスの間違いを発見した場合や、より適切な訳語に修正する必要が生じた場合など、特許文書の翻訳時に訳語を変更する必要は生じうる。

このような訳語を変更する最も単純な方法は、望ましい訳語が使われている翻訳例を対訳コーパスに追加する方法である。追加自体は簡単であるが、SMT の場合は、対訳を一文

[11] Masaru Fuji, A. Fujita, M. Utiyama, E. Sumita, Y. Matsumoto: Patent Claim Translation based on Sublanguage-specific Sentence Structure. MT Summit, 2015.

対追加したからといって期待される訳語が出力されるようになるかは学習そして翻訳をしてみないと分からない。未知語の場合には、一文対追加しただけで期待される訳が出力されるようになる可能性もあるが、既に別の訳語で翻訳がなされている場合には、新たに登録した一文対だけでは訳が変わらない場合もある。また、SMT の学習は、対訳コーパスのサイズが大きくなるにしたがって長い時間がかかるようになるため、対訳事例を追加してから結果を確認できるまで長い時間を要する。以上の点から、単純に対訳を追加する手法はカスタマイズの観点からは適切とは言いがたい。

このような問題を解決するため、以下のようなカスタマイズ機能を有する SMT エンジンが開発されている[12]。

・特定の見出しに対して訳語を指定する。
・翻訳に使ってほしくないフレーズをブラックリストとして指定する。
・出力結果の訳文を書き換える。

なお、オープンソフトの統計翻訳のツールとして広く使われている Moses[13]では、入力文の単語列を XML 形式で指定する動作モードにおいて、訳語を XML タグの属性値として指定することができるようになっている。

1.1.7　ドメイン適応の容易性

機械翻訳におけるドメイン適応とは、汎用の翻訳エンジンとして開発されたものを、特定の分野においてより高い翻訳精度が得られるように改良する行為のことである[14]。SMT による特許翻訳という観点で考えると、複数の技術分野をカバーする大量の対訳コーパスがあることを前提として、さらにその上で特定の技術分野に特化した小規模の別の対訳コーパスが存在する時に、それらを組み合わせて当該分野で翻訳精度が高い SMT エンジンを構築するのがその一例である。特許公報を全て翻訳する目的で翻訳システムを構築する場合にはあまり活用する機会はないかもしれないが、特定の企業が自社製品に関連する技術分野で対訳コーパスを収集して独自の翻訳システムを構築したいという場合には、ドメイン適応を行う意味がある。

文献 [12] においては、適応に利用できる対訳文が非常に少ない場合（例えば 1～1,000 文）と、ある程度大きい場合（10～20 万文）の 2 種類の場合についてドメイン適応のやり方を提案している。対訳文が少量の場合には単語アライメントの正確性が翻訳モデルを作

[12] 内山将夫,隅田英一郎:機械翻訳のドメイン適応とカスタマイズの事例. 言語処理学会第 22 回年次大会発表論文集, pp.529-532, 2016.
[13] Moses, http://www.statmt.org/moses/（最終検索日:2016 年 6 月 30 日）
[14] 特定の技術分野向けにチューニングする以外にも、利用者が翻訳したい文書の種類（例えば特許など科学技術文献の他、ニュースやメール、マニュアル、Web ページなど）に応じてより適切な翻訳結果が得られるようにチューニングすることも広い意味ではドメイン適応とみなすこともできるが、本調査は特許翻訳用の機械翻訳システムに関するものであるので文書の種類の多様性については考慮する必要がないと考え、ここでは技術分野に関する視点でのみ整理する。

る際のボトルネックになるため、単語アライメントにおける対訳単語間の確率推定において、当該対訳コーパスから推定した確率と大規模データから推定した確率を補完した確率を活用することを提案している。また、対訳文がある程度大きい場合の手法としては、大規模コーパスからドメインに類似する文を選択したり、確率モデルを適応したり、フレーズテーブルを組み合わせる手法など様々な手法が研究されているが、上記文献においては、翻訳モデルを線形補完する方法と、２つのサイズの異なる対訳コーパスから求めたフレーズペアの出現回数を合計して得られたフレーズペアから翻訳モデルを作成する方法の２つを比較している。その結果、後者の手法が最も翻訳精度が高かったと報告されている。

1.1.8　新たな言語対の追加容易性

既に述べているように、対訳コーパスを用意すればどのような言語であっても、ある程度の翻訳精度の翻訳システムを構築できるのが SMT の特徴である。どの程度の翻訳精度が必要であるかは、翻訳システムの利用目的や原言語に対する利用者の知識レベルにも依存するが、技術分野が広範囲に渡る特許の分野で実用的な翻訳精度を望むとすれば、最低でも 100 万文対程度の対訳コーパスは用意する必要があると考えられる。もしこの規模のコーパスがパテントファミリーの特許文書から自動構築できるのであれば、新たな言語対の追加も容易であると言える。しかし、そういった規模のコーパスが構築できるパテントファミリーが存在しない場合には、文同士の対応付けは難しいが類似した内容について記載した文書を用いてそれらから抽出された対訳候補を人手でチェックするか、人手により翻訳そのものを行って対訳コーパスを準備する必要がある。後者の場合は、コストが大きくなるので現実的ではなく、結局のところ特許分野ではパテントファミリーの文書が大量に存在するかどうか、具体的には１文書から 100 文対程度が対訳として抽出できると仮定して、100 万文対の対訳コーパスを作るにはパテントファミリーが１万文献以上存在するか否かで実現性が決まると考えることができる。

また、関心のある特定の言語対で大規模な対訳コーパスを構築できない場合には、1.5 で述べるピボット言語を媒介とした翻訳で実現するといったアプローチを検討する必要がある。

1.1.9　ノイズに対する頑健性

SMT は、入力文をフレーズに分割し、その単位で対応する目標言語のフレーズに置き換えることで翻訳を行っている。ここでいうフレーズとは、言語学的に意味のある単位である名詞句や動詞句といったものではなく、あくまで翻訳する上で固定的に取り扱える１語以上の単語の塊でしかない。そのため、原文に文法上の誤りが含まれる場合でも、翻訳実行時に最も目標言語として尤もらしい訳文に変換できるフレーズが見つけ出され、それなりの訳文が出力される。その意味で、全ての言語現象を規則によってとらえようとする

RBMT と比べて、SMT は原文の文法誤りなどのノイズに対して強いという特徴がある。

ただし、SMT でも PRE-ORDERING 技術による語順の変更を行う場合には、最初に原文の構文解析を行って構文構造を認識し、その構造の情報をもとに語順を変換するため、PRE-ORDERING を行わない SMT と比べると、ノイズに対する頑健性は低下していると考えられる。

1.2 ルールベース翻訳（RBMT）

1.2.1 RBMT の概要

ルールベース翻訳（RBMT）は、辞書と文法規則を主たる情報として翻訳処理を行う方式である。これは、原文の構文構造（構文解析）もしくは意味構造を認識（意味解析）し、次に各形態素を翻訳（語彙トランスファ）し、目標言語の構文構造に変換（構文トランスファ）し、意味構造に変換（構文構造生成）し、最後に訳文を生成（形態素構造生成）する方法でトランスファ方式とも呼ばれる。図１．２．１に RBMT の動作例を示す。

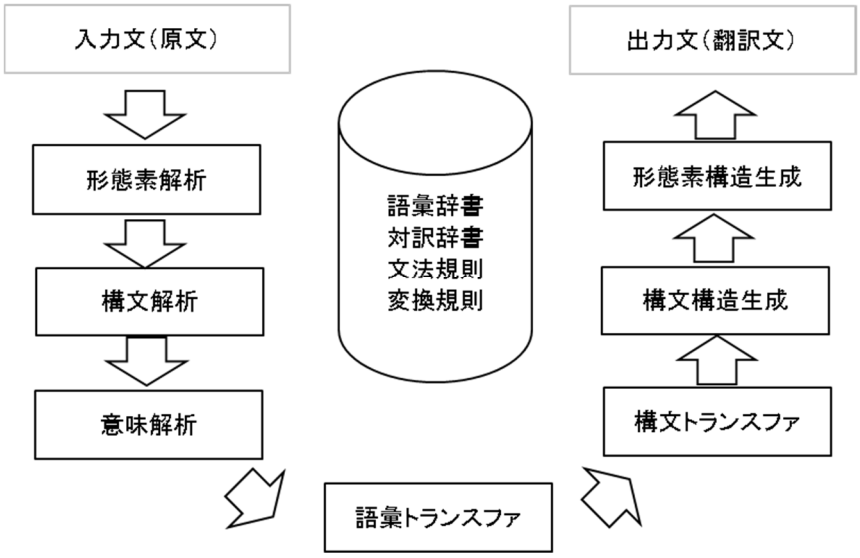


図１．２．１　RBMT の動作例

RBMT における各処理過程においては、原言語の単語の見出しや品詞を収録した語彙辞書や、原言語の各単語に対する訳語を定義した対訳辞書、原文の構文構造を解析するための文法規則、原言語の構文構造や意味構造を目標言語での対応する構造に変換するための変換規則、目標言語での語順や派生形を決定する文法規則などが利用される。

RBMT に関しては長年の研究開発を経て方式的には既に確立されており、近年新たに方式上の革新がもたらされたという報告は見受けられないが、近年利用可能になってきた大規模な言語資源や統計的な手法の知見を活用し、精度改善が図られている。すなわち、RBMT で利用される各種情報は従来人手で構築されてきたが、これを大規模な言語資源から抽出

して活用する試みがなされており、構文解析や語の訳し分けの精度向上に利用されている。中でも、RBMT において最も基本的な情報である翻訳辞書の構築を、対訳コーパスを利用して効率的に行う研究が精力的に行われており、成果を上げている[15] [16] [17]。近年では、SMT における学習の結果得られるフレーズテーブルを辞書作成に利用する手法も提案されている[18] [19] [20]。また、インターネット上の多言語の百科事典であるウィキペディア（Wikipedia）を利用して辞書作成を行う手法の研究もなされている[21] [22]。

1.2.2 RBMT の一般的なメリット、デメリット

RBMT を他の翻訳方式、特に SMT と比較した場合のメリット、デメリットは以下のとおりである。

(1) RBMT のメリット

① 翻訳対象規則が小さい場合のシステム開発が容易であること

RBMT は、翻訳対象となっている言語の構造や、対象言語間の関係を、比較的少数の規則で抽象化するため、ある程度の翻訳精度を有するシステムを容易に構築できる。

② 訳語の指定と統制が容易であること

SMT は原文に出現しない語が訳文に現れることや、原文に書かれている内容語が訳出されないといった問題が生ずるのに対し、RBMT は辞書で訳語が指定できるため、文書内の訳語を統制しやすい。また、SMT は、原文の文脈が訳語の選定に影響するので、一つの原文文書内で複数回出現する単語がある場合、それぞれの訳語が変化することがあるが、RBMT は辞書で訳語が指定されているので、訳語を統一できる。

③ 翻訳速度が高速なこと

経験上、RBMT は SMT と比較して翻訳速度が２桁程度速い。また、処理に必要なメモリ量も２桁程度少ない（SMT の翻訳速度や必要なメモリ量は、使用する翻訳モデルや言語モデルのサイズに依存する）。

---

15 Kupiec, J.: An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora, Proc. of the 31th Annual Meeting of the ACL, pp.17-22（1993）.

16 山本由紀雄,坂本仁:対訳コーパスを用いた専門用語対訳辞書の作成,情報処理学会研究報告,NL94-12（1993）.

17 熊野明,平川秀樹:対訳文書からの機械翻訳専門用語辞書作成. 情報処理学会論文誌, Vol.35 No.11, 1994.

18 森下洋平,梁冰,宇津呂武仁,山本幹雄:フレーズテーブルおよび既存対訳辞書を用いた専門用語の訳語推定. 電子情報通信学会論文誌, D, Vol.J93—D, No.11, pp. 2525-2537, 2010.

19 特許庁 平成 23（2011）年度特許文献の機械翻訳のための辞書データ整備に関する調査

20 安田圭志,隅田英一郎:日中特許対訳コーパスを用いた対訳辞書の自動構築. 言語処理学会 第 19 回年次大会 発表論文集,pp.306-309, 2013.

21 Maike Erdmann, Kotaro Nakayama, Takahiro Hara and Shojiro Nishio: Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia. The Information Processing Society of Japan（IPSJ） Journal（Jul. 2008）, 2008.

22 岡崎直観,劉瀟,綱川隆司,辻井潤一: Wikipedia からの対訳辞書候補の抽出. The 23rd Annual Conference of the Japanese Society for Artificial Intelligence, 2009.

④　翻訳結果の改良が容易であること

RBMT は、本来期待される訳文が得られなかった場合に、その原因を分析し所望の訳文が出力されるようにシステムを改良しやすい。例えば、特許の翻訳において特定の専門用語が適切に翻訳されていない場合には、その用語を辞書に登録すれば多くの場合適切に翻訳されるようになるが、その登録に要する時間は比較的短時間で済む。SMT の場合には、例文を対訳コーパスに追加した上で再学習を行う必要があるため、RBMT のように短時間でその効果を確認することができない。

(2) RBMT のデメリット

①　訳文が自然とはいいにくいこと

RBMT は SMT や EBMT と比較して、訳文が自然とはいいにくい。

②　システムの多言語化が難しいこと

RBMT は、文法規則や辞書の記述方法がシステムにより異なるため、SMT に比べ、システムの多言語化が困難である。

③　翻訳規則の変更に時間がかかること

RBMT は、規則を追加する際の副作用に注意する必要がある。すなわち、既存の規則との整合性を開発者が把握した上で追加しないと、予期しない翻訳結果が出力されるようになり、副作用の発見とその原因分析や対策に時間がかかる。

1.2.3　特許分野での適用上の課題

特許文献は一般文書と比べ、長文が多いため、そういった長文をどう扱うかは、特許情報の機械翻訳における課題である。ただ、請求項のように、長いが特定のパターンで表現されることが多いものについては、原文を予め翻訳しやすい形に修正してから翻訳するというアプローチを取ることができる。RBMT では、こういった前編集処理を、翻訳における原文解析と同様に規則に基づく処理で行いやすく、既存の製品の中にはこのような機能が提供されている製品もある[23]。

1.2.4　多言語化の現状

特許文献の機械翻訳において、RBMT 方式の英日・中日・韓日翻訳は既に実用化されている。一方、ベトナム語、タイ語、インドネシア語と日本語の間の機械翻訳システムにおいて、RBMT は実用化していない。実用化の有無が分かれた理由として翻訳ニーズの有無が考

[23] 鈴木博和,熊野明:特許文書用前編集機能を備えた機械翻訳システム. 情報処理学会第 63 回国大会全（2001）.

えられる。つまり、英語、日本語、中国語、韓国語は以前から翻訳のニーズがあったため、対訳コーパスや辞書の開発がなされていたのに対してベトナム語、タイ語、インドネシア語の翻訳ニーズが高まっていたのは近年になってからであるため、翻訳の研究が成熟していないことが原因として挙げられる。

1.2.5 翻訳精度の現状、精度向上に向けた取り組み

(1) 形態素解析

形態素解析の役割は、主に以下の３つがある。

① 日本語のような分かち書きされていない言語（日本語など）対して行う単語の区切りを決定する処理
② 活用する語に対して行う、原形及び語幹を同定する処理
③ 品詞を割り当てる処理

統計的手法が実用化されるまでは、これらの処理を全て言語的知識に基づいて行っていたが、近年はタグ付きコーパスを用いた学習による統計的手法が主流になってきている[24][25]。より詳しくは 3 節を参照のこと。これにより RBMT においても、形態素解析は統計的な手法を用いて行うことが可能になっており、大規模なコーパスで学習することで、形態素解析の精度を従来よりも向上させることが可能な環境が整ってきた。ただし、商用の RBMT システムは、処理全体がブラックボックスとなっているため、形態素解析をどのように行っているかをユーザが推測することは難しい。

(2) 構文解析／意味解析

形態素解析と同様、構文解析や意味解析においても、構文・意味情報が付与されたタグ付きコーパスから抽出した確率値をもとに、構文的曖昧性や意味的曖昧性を解消する統計的な手法が実用化されている。これにより、英語のみならず日本語や中国語においても、従来の RBMT と比較して高い精度の解析が可能となってきており[26][27][28][29][30]、構文解析や

[24] ChaSen -- 形態素解析器, http://chasen-legacy.osdn.jp/（最終検索日:2016 年 6 月 30 日）
[25] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://taku910.github.io/mecab/（最終検索日:2016 年 6 月 30 日）
[26] Dan Klein, Christopher D. Manning: Parsing with Treebank Grammars: Empirical Bounds, Theoretical Models, and the Structure of the Penn Treebank. Proceedings of 39th Annual Meeting of the Association for Computational Linguistics, 2001.
[27] 白井清昭,乾健太郎,徳永健伸,田中穂積:統計的構文解析における構文的統計情報と語彙的統計情報の統合について.自然言語処理, Vol.14, No.4, pp.67-81, 2007.
[28] 河原大輔,黒橋禎夫:自動構築した増規模格フレームに基づく構文格解析の統合的確率モデル.自然言語処理, Vol.14, No.4, pp.67-81, 2007.
[29] 工藤拓,松本裕治:チャンキングの段階適用による日本語係り受け解析.情報処理学会論文誌, Vol.43, No.6, pp.1834-1842, 2002.
[30] Roger Levy, Christopher Manning: Is it harder to parse Chinese, or the Chinese Treebank? Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, 2003.

意味解析を SMT に適用して、原言語の語順変換の精度を上げ、翻訳精度を向上させた例もある[31]。

(3) 構造変換並びに訳語選択

トランスファ方式の RBMT において、原言語と目標言語間の構文構造若しくは意味構造をブリッジするための変換規則や訳語選択規則を、対訳コーパスから獲得する研究も行われている。例えば、参考文献[32]では、単言語コーパスに含まれる言語現象をもとに、人間が効率的に翻訳規則を拡充する方法が提案されている。また、文献[33]では対訳文に対して句のアライメントを行う手法が提案されているが、対応づけられた原言語と目標言語の句から翻訳規則を作ることができる。また、対訳コーパスから機能表現の翻訳パターンや、表現を部分的に変数と置き換えた翻訳規則を獲得する手法も提案されている[34] [35]。

1.2.6 カスタマイズの容易性

1.2.2 で述べたように、RBMT では、所望の訳文が得られなかった場合に、システムを改良しやすい。ただし、翻訳エンジンのベンダーであれば、どのような改良も可能であるが、ユーザが行えるカスタマイズ手段は限られている。

カスタマイズにおいて最も基本的な方法は、専門用語辞書の選択とユーザ辞書への語彙の追加である。専門用語辞書の選択とは、ベンダーが用意した複数の専門用語辞書を、自分が翻訳したい専門分野に応じて選択することである。特許の翻訳においては、専門分野が多岐に渡るため、同一の見出しであっても、専門分野によって異なる訳語を選択する必要があるためである。専門用語辞書に収録されていない用語については、見出しとともに訳語をユーザ辞書に登録することで対応することができる。これにより多くの専門用語は対処可能であり、市販の RBMT のソフトにおいては、ほとんどすべてのソフトにて登録機能が提供されている。

一方、動詞の訳し分けのように、主語や目的語との組み合わせにより異なる訳語を使い分ける必要がある場合には、そういった訳し分け規則を登録できる機能が必要となる。このような機能を提供しているソフトも存在するが、SMT のように対訳文を追加するのと比べると、作業の専門性が高くなるため、特許分野のように数多くの専門分野に渡って大規模に登録するのには向いていないと言える。

なお、以上述べたような形のカスタマイズは、基本的にはそれぞれの辞書にデータを追

---

31 Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, Masaaki Nagata, Xianchao Wu, Takuya Matsuzaki, Jun'ichi Tsujii NTT-UT Statistical Machine Translation in NTCIR-9 PatentMT http://cl.naist.jp/~kevinduh/papers/sudoh11ntcir.pdf

32 山田節夫,今村賢治,山本和英:コーパスを利用した効率的な翻訳規則の拡充:自然言語処理, Vol.13, No.2, pp.3-18, 2006.

33 荒牧英治,黒橋禎夫,佐藤理史,渡辺日出雄:用例ベース翻訳のための対訳文の句アラインメント. 自然言語処理, Vol.10, No.5, pp.75-92, 2003.

34 黒橋禎夫,中澤敏明:対訳コーパスからの機能表現翻訳パターンの自動学習. Japio 2008 YEAR BOOK.

35 寺島涼,越前谷博,荒木健治:フレーズベース統計翻訳と対訳文一般化による翻訳規則の統合. 言語処理学会 第 17 回年次大会 発表論文集, pp.175-177, 2011.

加するだけでよく、SMT のように翻訳モデルの再学習を行わなくてもただちに翻訳結果に反映される。そのため、登録した知識の効果がすぐに確認でき、個別の誤訳に対するカスタマイズ作業は比較的短時間で行うことができる。

1.2.7　ドメイン適応の容易性

SMT においては、翻訳モデルの訓練に使用する訓練コーパスを適応させたいドメインに応じて準備することで、どのような粒度で分割したドメインであっても、それらドメインごとに訓練さえ行えば、各ドメインで適切に訓練された翻訳エンジンを用意することが可能となる。そのため、現時点ですでに大規模な特許対訳コーパスが構築されている日本語と英中韓３言語間については、学習に要する計算資源の問題がなければ、必要な粒度でドメインごとに学習を行うことが可能な状況になってきている。

これに対し、RBMT の場合には、特定技術分野の翻訳においてどういった翻訳辞書を使用するのかを決定するとともに、翻訳されない用語が見つかった場合にユーザ辞書に登録していく作業がドメイン適応となる。一般には翻訳エンジンのベンダーから複数の専門用語辞書が提供されているが、全ての辞書を同時に使えるとは限らず、また、利用可能な最大数の専門用語辞書を使用したからといって翻訳精度が上がるわけではない。そのため、どのような辞書をどのような優先順位で組み合わせて使用すると翻訳精度が向上するかを決定する必要がある。

通常、機械翻訳は、Web 上の翻訳サービスのように分野を限定しない形で使用されるか、ユーザ自身が関心のある特定の技術分野に限定した形でドメイン適応を実施することが多いため、特許のように技術分野が多岐に渡る文献を翻訳する際のドメインの適応方法に関する研究は少ない[36]。

1.2.8　新たな言語対の追加容易性

RBMT では、翻訳システムで使用する知識を辞書、文法の形で整備する必要がある。1.2.1 で述べたように、近年、そのような知識を大規模な言語資源を利用して開発できるようになってきているため、一から人手で開発していた時代に比べると、開発期間の短縮や開発コストの低減が可能となっているが、対訳コーパスを用意すればエンジンの開発が可能な SMT と比べると、新たな言語対の追加にあたっての敷居は高い。すなわち、原文を形態素に分割する形態素解析器や、原文の構文構造を解析し事前に目標言語の語順に近い形に形態素列の語順を入れ替える PRE-ORDERING のための構文解析器は SMT とほぼ共通であるとしても、RBMT では、変換処理のための変換規則や、変換処理によって得られた目標言語の意味構造から語順を決定する構文生成規則などの文法を開発する必要がある。それらの開発には、言語処理の知識を有する技術者や研究者が必要であり、そのための開発期

[36] Satoshi Sonoh, Satoshi Kinoshita, Hiroyuki Tanaka and Satoshi Kamatani： Toshiba MT System Description for the WAT2014 Workshop. In Proc. of the 1st Workshop on Asian Translation, 2014.

間やコストが必要となるため、新たな言語対の追加は SMT よりも難易度が高くなると考えられる。参考文献[37]には、定量的なコスト比較の記載はないが、『コーパスベース翻訳技術により、従来の「ルールベース翻訳技術」の抱える開発コストや多言語展開への困難さといった問題点が改善された』との記載がある。

1.2.9 ノイズに対する頑健性

RBMT は、翻訳対象となる言語の現象を文法規則という抽象化されたレベルで認識して処理するため、本質的に翻訳システムに登録された規則を使って適切に解析できた範囲でしか正しい翻訳ができない。そのため、SMT と比べると、未知語や原文の文法誤りに対する頑健性がない。例えば、未知語の場合、正しく品詞が推定できるかが鍵となる。英語の場合、ly で終わっていれば副詞の可能性、ed で終わっていれば動詞の可能性があるため、辞書に登録されていない未知語であっても名詞の可能性以外にそれらの品詞の可能性を仮説として設定して原文を解析し、解析に成功すればその仮説を採用するといった処理を行うことができる。しかし、そういった推定が適切に行えない場合、構文解析が失敗し、全体として理解不能な訳文を出力してしまう場合がある。同様に、英語であれば冠詞が連続するといった文法の誤りがあると、やはり構文解析に失敗し、文全体として読むに堪えない訳文が出力されてしまう場合がある。そこで、このような文法誤りがある場合でも、文全体の構文構造としてある程度許容できるような解析結果が得られるようにするための構文解析の手法が提案されており、機械翻訳への適用で効果も確認されている[38]。

1.3 用例に基づく翻訳（EBMT）

1.3.1 EBMT の概要

用例に基づく機械翻訳（Example-Based Machine Translation, EBMT）とは、過去の翻訳例を模倣して新たな文を翻訳する手法で、アナロジーに基づく機械翻訳とも呼ばれる[39]。以下では手法の概要を、文献[40]を参考に説明する。

RBMT では、解析・変換・生成の情報をもとに翻訳を行うが、言語現象は多様であり、全ての場合を尽くすことはほぼ不可能である。例えば、日本語で「A の B」（A, B は共に名詞）という表現を英語に翻訳する場合を考えてみる[41]。表１．３．１-１の一つ目の例のように、多くの場合は“B of A”とすれば良い。しかし、その他の例のように、“B of

37 ATR プレスリリース資料, http://www.atr.jp/topics/press_080325_final.pdf（最終検索日:2016 年 6 月 30 日）

38 小田悠介, Graham Neubig,波多腰優斗,Sakriani Sakti,戸田智基,中村哲:解析失敗の発生しにくい PCFG-LA 句構造構文解析. 言語処理学会 第 21 回年次大会 講演論文集, 2015.

39 Makoto Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. In Proc. Of the International NATO Symposium on Artificial and Human Intelligence, pages 173–180, New York, NY, USA, 1984. Elsevier North-Holland, Inc.

40 渡辺太郎,今村賢治,賀沢秀人,Graham Neubig, and 中澤敏明. 機械翻訳. コロナ社, 2014. 1.3 用例に基づく機械翻訳

41 Eiichiro Sumita, Hitoshi Iida, and Hideo Kohyama. Translating with examples: A new approach to machine translation. In TMI, pages 203–212, 6 1990.

A”という翻訳が不自然、又は不適切な場合も多く存在する。このように、短い名詞句であってもその翻訳パターンは多様であり、さらに大きな単位である節や文まで考慮すると、その全てについて翻訳規則を書き下すことが困難であることは想像に難くない。また、翻訳規則の追加が他の規則と衝突することのないように注意を払う必要もある。このように、RBMT はシステム構築の労力が大きい。

表1．3．1-1　日本語名詞句「A の B」の訳し分け

A の B	英語訳
八日　の　午後	the afternoon of the 8th
会議　の　参加料	the application fee for the conference
三つ　の　ホテル	three hotels
京都　の　ホテル	hotels in Kyoto
太郎　の　母親	Taro’s mother
崖　の　上	on the cliff

これに対して、EBMT では、文の解析・変換・生成の規則を記述する代わりに、過去の翻訳例（=トレーニング対訳コーパス）をもとに翻訳を行う。例えば、「パリのホテル」を翻訳したい場合は、これとよく似た「京都のホテル」の例文を使い、京都の部分をパリに置き換えることで、“hotels in Paris”という訳が得られる。つまり、参考となる翻訳例さえあれば、各言語の深い知識は不要である[42]。また、システムが翻訳できる文のパターン、つまり翻訳能力は、基本的に対訳文数と相関する。逆に言えば、対訳文を増強するだけで、簡単に翻訳能力を向上することができ、システム全体のメンテナンスも容易である。

EBMT では、対訳コーパスが必須であるが、これは SMT も同じであり、両者はコーパスに基づく機械翻訳（Corpus-Based Machine Translation,CBMT）あるいはデータに基づく機械翻訳（Data-Driven Machine Translation）と呼ばれることもある。コーパスに基づく機械翻訳では、各対訳文内の単語又は句の対応関係が与えられた上で、そこから翻訳知識の学習を行う。しかし、翻訳知識の学習に用いる対訳文は数百万にものぼるため、大量のコーパスに対して人手で対応関係を付与することは事実上不可能であり、自動的に対応関係を推定する必要がある。

(1) 事例ベース推論としての機械翻訳

用例に基づく機械翻訳は、しばしば事例ベース推論（Case-Based Reasoning, CBR）の機械翻訳における具体化と言われる。事例ベース推論とは、過去の似た事例の解法に基づ

[42] ただし、深い知識は不要とはいえ、翻訳対象言語の文を解析して依存構造を作り出す解析技術は必要なので、翻訳例だけを用意すれば翻訳ができるというわけではない。

いて新たな問題を解く方法又はその過程のことであり、人間の問題解決手法としても日常的に用いられるものである。人間の取る行動や意思決定は、多くの場合、規則として記述されたものではなく、過去の経験によるところが大きい。このように事例ベース推論は「問題解決方法を規則という形で明確に定義することは困難だが，解決事例は数多く容易に得られる」場合に適している。例えば、裁判における判例の利用、医療分野での症状からの病名診断などである。機械翻訳においては RBMT のように規則を書き下すことは大変な労力がかかるが、EBMT と SMT は対訳コーパスという形で数多くの事例を得ることができる。

用例に基づく機械翻訳は、大きくわけて以下の２つの処理からなる。

・入力文と類似する用例を対訳コーパスから検索
・得られた用例を入力文に合わせて修正、又は複数の用例を結合

最大の特徴は、事前に翻訳の知識や確率モデルなどを仮定せず、入力文と用例との類似度計算などが翻訳時に行われるという点である。それゆえ、この類似度をどのように計算するかが最も重要である。また、入力文との異なりが数単語しかないような用例が見つかる場合はそう多くはないため、複数の用例を組み合わせて利用する必要がある。以下ではこの二点について詳しく述べる。

(2) 用例の検索と修正

日本語の「掛ける」には様々な意味があり、日本語語彙大系[43]には 100 以上もの意味が定められている。その一部を表１．３．１-２に示す。このように多義性のある語を英語に翻訳する際には、訳語選択を適切に行う必要がある。RBMT では、格パターン対応規則をあらかじめ用意するなどして対応するが、EBMT では、入力文と原言語側で類似度の高い用例を選択することで対応する。

表１．３．１-２　日本語「掛ける」の訳し分け

「[名詞]を掛ける」の文の例	英語訳
お金を掛ける	spend money
コートを掛ける	hang a coat on
CD を掛ける	play a CD
眼鏡を掛ける	wear glasses
目覚ましを掛ける	set an alarm clock

翻訳の際、完全に一致するものが用例にある場合、その訳をそのまま出力する。完全に

[43] NTT コミュニケーション科学研究所 監修 池原悟他 日本語語彙大系 岩波書店 1997.

一致するものがない場合は、極力「似た文」を探すが、これには、入力文とそれぞれの用例との類似度を計算する必要があり、その尺度は様々である。単純には文間の編集距離（edit distance）を使うことが考えられる。編集距離は、ある文字列から別の文字列に変換する際に必要となる、文字や単語の削除・挿入・置換の操作回数により定義されるが、この例ではどの文とも１単語しか異ならず区別がつかない。なお，翻訳会社などで広く使われている TRADOS のような翻訳メモリ（translation memory）管理システムは、編集距離をもとに入力文と翻訳メモリとの距離を計算しているが、これらのシステムは似た文を翻訳作業者に提示するだけであり、最終的な翻訳は人間が行う必要があるため翻訳システムとは本質的に異なる。

より良い類似度の指標として、異なる部分の単語の類似性を考慮することが考えられる。単語の意味的な近さは、シソーラス（thesaurus）などの人手で整備されたリソースを用いたり、分布類似度（distributional similarity）のような自動獲得知識を使ったりするなど様々である。シソーラスとは，言葉を意味の単位で分類・配列し、意味ごとに同義語・上位語・下位語などの階層構造を用いて表したものであり、この階層構造を辿ることによって語と語の距離を定義することが可能である[44]。一方、分布類似度は、「似た語は似た文脈で出現する」という分布仮説に基づいて計算される語の類似度で、単言語の大規模コーパスさえあれば計算できるため、人手による整備コストが限りなく小さく、また新語などに対しても頑健である。

これらの方法により、「レコード」と「お金」「コート」「ＣＤ」「眼鏡」「目覚まし」との類似度を計算することで、意味的に最も近いものは音楽に関係する「ＣＤ」であり、入力文と最も似ている用例は「ＣＤを掛ける↔play a CD」であることが分かる。あとは、用例と入力で異なっている部分を、対訳辞書を用いるなどして正しく修正し、最終的に“play a record”という訳が得られる。RBMT では動詞の訳し分けは構造の変換時に行われたが，EBMT では、ここで述べたように用例の検索によって行われる。

入力文と用例の違いが複数箇所あっても、基本的には同じ方法で翻訳が可能である。例えば、表１．３．１-１の対訳文を用例として、「箱根の旅館」を翻訳することを考える。この場合、「箱根」及び「旅館」と、それぞれの場所にある語との類似度をそれぞれ計算し、類似度の和が最も大きい文を選択すれば良い。「旅館」と類似度の高い「ホテル」を含む文は「三つのホテル」と「京都のホテル」の２つがあるが、「箱根」と類似度が高いのは「京都」であるので、「京都のホテル↔hotels in Kyoto」の例文が選択される。最後に「京都」と「ホテル」をそれぞれ「箱根」と「旅館」に修正すれば、“guesthouses in Hakone”という翻訳が得られる。

[44] フリーのリソースとして NICT より日本語 WordNet が配布されている: http://nlpwww.nict.go.jp/wn-ja/（最終検索日:2016 年 7 月 14 日）

(3) 複数用例の利用

入力文と異なる部分が数単語だけである用例が常に見つかるならば、これまで述べた方法で問題はないが、一般的に対訳コーパスのカバー率は高くはなく、用例を少し修正するだけで翻訳が得られることはほとんどない。そこで、より多くのパターンの入力文に対応するために、複数の用例から部分的な訳を取り出し、それらを組み合わせることによって翻訳を生成する。

「彼女は機械翻訳の教科書を読んだ」という文の翻訳を考える。また、用例として

・彼は本を読んだ ⇔ He read a book
・私は昨日教科書を買った ⇔ I bought a textbook yesterday
・彼女は機械翻訳の論文を印刷した ⇔ She printed out a paper on machine translation

という３つの対訳文が与えられているとする。例えば、下線を引いた部分を組み合わせることで、“She read a textbook on machine translation”という訳を生成することができる。

では、これを計算機上で実現するにはどうしたらよいだろうか。ここで入力文及び用例は、全て単語依存構造木で与えられるものとする。上記入力文及び３つの用例の単語依存構造木を図１．３．１-１に示す。

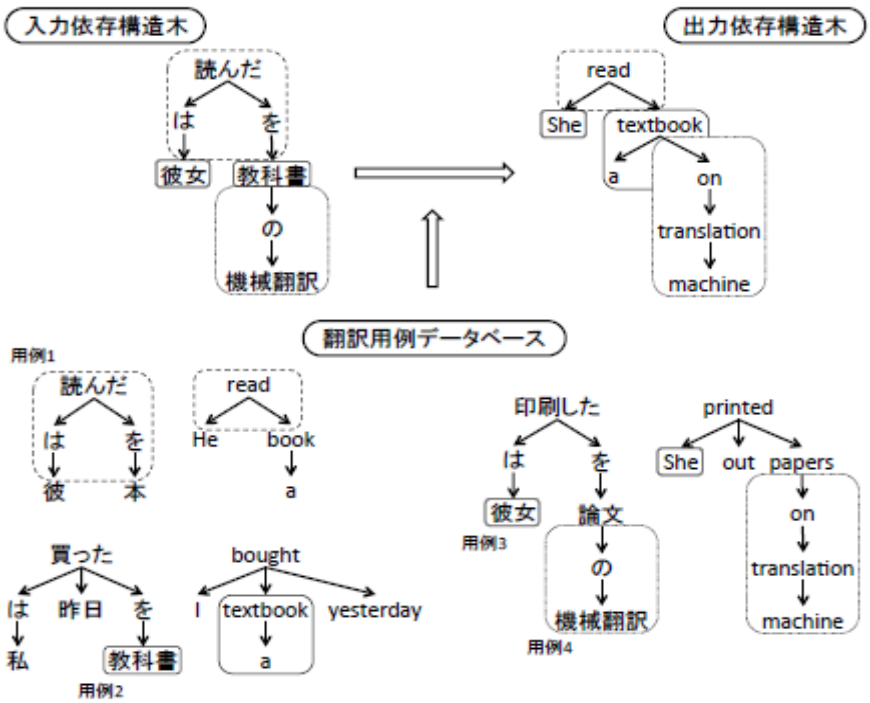


図１．３．１-１　各文の依存構造木と、用例に基づく機械翻訳の動作例

ここまでの説明では、対訳文の一つ一つを用例と呼んでいたが、複数用例を利用した翻訳を実現するために、用例の粒度を細かくする。具体的には、対訳文内の各単語対応及び、対訳文から単語対応を除いたものを用例と呼ぶこととする。ただし、用例は両言語において木構造上連続であるものとする。また、これまでとおり文全体も用例である。例えば、「彼は本を読んだ ⇔ He read a book」の用例からは、以下の用例が得られる。

・ 彼 ⇔ He
・ 本 ⇔ a book
・(X) は (Y) を読んだ ⇔ (X) read (Y)
・彼は (Y) を読んだ ⇔ He read (Y)
・(X) は本を読んだ ⇔ (X) read a book
・彼は本を読んだ ⇔ He read a book

他の対訳文からも同様に、可能な用例を全て作り出す。ここで問題となるのは、日本語の助詞や英語の冠詞など、相手言語に対応する語が存在しない単語の扱いであるが、現在のところ、決定的な方法は提案されておらず、今後の研究課題となっている。

用例の準備ができたら、次は入力文を翻訳するのに必要な用例の検索を行う。これは、入力文の全ての部分木について、用例の中からそれぞれの部分木をカバーするものを探すことで実現できる。入力文「彼女は機械翻訳の教科書を読んだ」に対しては、以下の４つの用例を発見することができる。

例１　(X) は (Y) を読んだ⇔ (X) read (Y)
例２　教科書⇔a textbook
例３　彼女⇔ She
例４　機械翻訳の (Z) ⇔ (Z) on machine translation

あとは、翻訳に利用可能な用例を組み合わせることで、最終的な翻訳を得ることができる。入力依存構造木で最も親側のノード、つまり入力文の根（root）に対応する用例を起点とし、順に子方向の用例を結合していく。例の入力文では、根は「読んだ」であるので、対応する例１が起点となり、これに他の用例を結合していく。

用例同士を結合する際には、用例本体のすぐ外側にある単語、つまり例１の (X) や (Y) にあたる部分を手がかりとすることができる。例１にある手がかりは、全て自分の子の位置にある手がかりであるので、これを子方向の手がかりと呼ぶことにする。子方向の手がかりを利用し、(X) を「彼女⇔She」に置き換え、さらに (Y) を「教科書⇔a textbook」に置き換えることで、「彼女は教科書を読んだ⇔She read a textbook」という

訳が得られる。

最後に、「機械翻訳の⇔on machine translation」を組み合わせるが、例１にはこれ以上組み合わせの手がかりとなる情報がない。そこで今度は、例４にある手がかりを用いる。例４には、自分の親の位置にある手がかり（Z）がある。これを親方向の手がかりと呼ぶ。入力文において、「機械翻訳の」の親にあたる単語は「教科書」であるため、その英語側の“ textbook ”が（Z）に相当する。逆に言えば、例４は例２の子の位置に結合すればよい。以上で全ての用例の結合が完了し、出力となる依存構造木が生成された。最後に、出力依存構造木の単語を順に読み出せば、最終的な翻訳である“ She read a text book on machine translation ”が得られる。

(4) 組み合わせの手がかり

複数の木構造から構成される用例を組み合わせる時、最も重要となるのはその手がかりの情報である。

子方向の手がかりは、他の用例を結合する場所に関して完全な情報を持っている。ここで言う完全な情報とは、手がかりとなるノードの兄弟にあたる語が存在したとしても、兄弟間の順序が既知であるので、単に手がかりとなる位置を他の用例に置き換えるだけで結合することができた。つまり兄弟間の語順を考える必要はない。

一方で、親方向の手がかりは、親の左側から結合するか、右側から結合するかといった結合の方向についての情報は持っているが、結合先に兄弟となる語があったとしても、それらと自らの用例との順序はわからない。正しい順番にするためには、言語的な制約を入れるか、言語モデルなどで判断する必要がある。最近では木構造の情報を利用することで、この問題を精度よく扱う方法も提案されている[45]。

なお、これらの用例の組み合わせは、木接合文法（Tree Adjoining Grammar, TAG）における置換（substitution）と接合（adjunction）操作を単語依存構造木に適用し、さらに２言語で同期的に行ったものと考えることができる。木接合文法は文脈自由文法に似た文法だが、書き換え規則が木構造を持っている点が特徴である。

(5) 最適な用例の組み合わせ

用例データベースが小規模の場合は、入力文を翻訳するための用例の組み合わせは一つしかない場合もあるが、用例データベースが大規模になると様々な用例が利用可能となり、結果として翻訳時に２種類の曖昧性が生じる。一つは、ある部分の翻訳において、様々な翻訳のパターンが存在するという曖昧性であり、もう一つは入力文の分割の曖昧性である。では、このような曖昧性の中で、最適な用例の組み合わせを求めるにはどのようにすればよいか。用例に基づく翻訳では，各用例にスコア付けをし、さらに翻訳に利用す

[45] John Richardson, Fabien Cromi`eres, Toshiaki Nakazawa, and Sadao Kurohashi. Flexible nonterminals for dependency tree-to-tree reordering. In NAACL, 2016.

る用例のスコアの和が最大となる分割を用いるという方法でこの問題を解決する。

用例のスコアは、様々な情報をもとに計算されるが、中でも重要な要素が三つある。一つ目は用例の大きさであり、より大きな用例に対してより大きなスコアを付与する。これは「大きな用例を少数用いる方が、翻訳が安定する」という経験則に基づいた指標である。二つ目は、用例の周囲の状況と、入力文においてその用例で翻訳される部分の周囲の状況との類似度である。これは1.3.1（2）で述べたような用例の周辺類似度、つまり、組み合わせの手がかりとなる部分の単語の類似度を用いればよい。三つ目は、翻訳の確からしさである。ある部分を翻訳するために得られた用例が複数ある場合、その訳が全て同じであるとは限らない。しかしながら、最も割合の高い訳を選べば、その訳が正しい可能性も高いと考え、検索された用例の中での訳の割合をスコア化し、割合の大きいものほど高いスコアを与える。

これらの主なスコア付け指標の他にも、手がかりとなる部分の品詞や、言語ごとに有効な指標などがある可能性もある。それらを柔軟に組み合わせることにより、適切な用例に高いスコアが与えられるようにする。用例のスコア付けが終われば、入力文を全てカバーする用例の組み合わせのうち、用例のスコアの和が最も大きくなる組み合わせを選ぶことにより、最適な用例の組み合わせを見つけることができる。

(6) 複雑な組み合わせ

これまでの例では、根となる用例を起点として、他の用例を子として順に結合していけば翻訳できたが、実際にはこのように簡単に翻訳できる文だけではない。最も頻繁に起こるのは主辞交代（head-switch）と呼ばれる現象である。これは、単言語でも起こる言語現象である。例えば、図１．３．１-２（a）のように、「赤い屋根の家」では「赤い」は「屋根」に係るが、「屋根が赤い家」では文の意味は同じであるが「屋根」が「赤い」に係るというように、係り受け関係が逆転する。これが二言語間でも起こり、例えば、図１．３．１-２（b）のように、「適用して成功した⇔successfully applied」だと、日本語では動詞である「成功した」が英語では副詞“successfully”となっており、依存関係だけでなく品詞も変化している。この用例を用いて、例えば、「実行して成功した」を翻訳することを考えると、入力文の根にあたる「成功した⇔successfully」に「実行して⇔executed」などの用例を結合するが、子としてではなく、親として結合する必要がある。

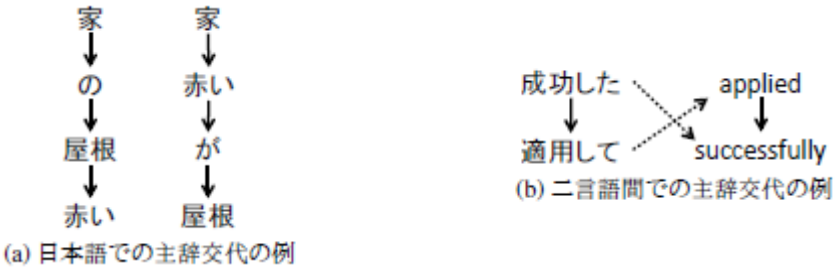


図１．３．１-２　主辞交代の例

(7) EBMT の現状

EBMT と SMT、RBMT を比較する形で、EBMT の現状を説明する。EBMT の特徴は、事前に翻訳の知識や確率モデルなどを仮定しないことだと述べたが、逆に言うと、扱う用例データベースのサイズが大きくなるにつれて適切な用例の選択が難しくなり、また、翻訳したい文が複雑になるにつれて正しい翻訳を生成することが難しくなるという問題がある。この問題に対して RBMT はルールごとに優先度などを設定して適切な翻訳を選択し、SMT は様々な素性（Moses の PBSMT では 14 個）を使って判断する。一方、初期の EBMT では RBMT のようなルールの優先順位がなく、素性の数も SMT ほど多くなかったため、他の手法に比べて適切な翻訳の選択が難しかった。

大量の対訳コーパスに基づく機械翻訳手法で複雑な文を正しく翻訳する試みは、現在ほとんどが SMT（若しくはごく最近ではニューラルネットワークを用いた手法）で行われている。一方 EBMT は、言語資源の乏しい言語対や、翻訳したい文のバリエーションがごく限られている場合に SMT よりうまく働くことが多いため、そのような条件下では EBMT が有効である。

京都大学の黒橋・河原研究室で研究開発している翻訳エンジン[46]は、おそらく日本で唯一の EBMT システムであるが、翻訳時には対数線形モデルにより様々な素性を考慮しており、PBSMT のデコード方法とほとんど同じである。このように、SMT と EBMT の本質的な違いはほとんどなくなっていると言っても過言ではない。

ただし、元の対訳文を可能な限りそのまま利用し、なるべく大きな用例を少数組み合わせて翻訳するという EBMT のアイデアは生きており、SMT のように１つ１つの翻訳ルールが小さくなっているものとは性質が異なる。また、SMT では翻訳ルールがフレーズテーブルの単純な文字列の対訳の形で保存されているため、元の対訳文とのつながりはなくなっており、翻訳結果と元の対訳文の語を対応付けることは容易ではなく、文法的な構造も失われやすい。一方、EBMT では翻訳時に適用された用例を追うことで翻訳結果から用例の元となった対訳文を特定することができ、誤った翻訳の原因の分析ができるなど、翻訳過程の

[46] John Richardson, Fabien Cromi`eres, Toshiaki Nakazawa, and Sadao Kurohashi. Kyotoebmt: An example-based dependency-to-dependency translation framework. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 79–84. Association for Computational Linguistics, 2014.

透明性が高いことが特徴である。これにより元の対訳文の誤りを発見し、修正するといったことも可能である。さらに、元の対訳文が保存されているということは、１つの対訳文を新たに追加すると、その対訳文と似た文は次からすぐに正確に翻訳できるようになるという利点もある。一方、SMT では対訳文がフレーズテーブルに文字列としてバラバラに保存されてしまうため、追加した対訳文と似た文を正確に翻訳できない場合がある。

また、現状の機械翻訳は、どの方式であれ翻訳文に誤りが含まれることは避けられない。そのため、大雑把に入力文の内容を把握したいときなどは十分有用であるが、翻訳してどこかに公表する場合や、新たな対訳文として利用したい場合などは、どうしても人の手を入れることが必要となる。その際、原文から翻訳するよりも、機械翻訳結果を後修正（post-edit）する方が効率的であるが、なぜその翻訳結果が出てきたのか原因が不明であると、修正作業者の精神的な負担が大きくなる。この点でも、EBMT は SMT と比べて有利であると言える。

(8) 今後の発展の可能性

ここまで述べたように、用例に基づく機械翻訳は、人間が脳内で行っているであろう翻訳のプロセスのモデル化の一つである。すなわち、与えられた入力文をいくつかの単位に分割し、過去に得た対訳知識から文構造や意味が類似するものを探し出し、単語や句の置き換えといった部分的な修正を施したり、複数の知識を組み合わせたりすることにより、所望の翻訳を得る。実際に、我々が外国語を学習するときにも、多くの例文とその訳を覚えることで、飛躍的に外国語の知識を拡充することができる。

上で述べたように、用例に基づく機械翻訳は、その翻訳過程の性質上、用例データベースと入力文とのドメインが近い場合、言い換えれば入力文と似たような文が用例データベースに多く含まれる場合に、特に有効な翻訳手法である。現実的には、そのような翻訳を行いたい状況は数多くある。例えば、バージョンアップされた製品のマニュアルを翻訳する際に、新バージョンで強化された機能についての説明文を翻訳するときには、前バージョンと表現や構文構造が類似した部分の対訳用例を用いることが考えられる。

一方で、用例に基づく機械翻訳の問題点として、用例選択の過程がヒューリスティックであるという点が挙げられる。用例のスコア付けの際に、どのような要素を考慮するかという点に柔軟性があることは利点ではあるが、実際にどの要素を選択するかや、各要素をどのようにスコア化し、さらにそれらをどのような割合で足し合わせるかなどについては、システムの開発者にゆだねられている。また、用例を一つ一つ個別の知識として蓄えているため、用例の数が大規模になればなるほど、翻訳時の用例の検索に時間がかかってしまう。しかしながら、これらの EBMT の問題は、近年の様々な研究により解決されつつある。

1.3.2 EBMT の一般的なメリットとデメリット

EBMT のメリットとデメリットを、SMT や RBMT などの他の翻訳方式と比較して説明する。

(1) EBMT のメリット

① 翻訳過程の透明性が高いこと

SMT では、翻訳ルールがフレーズテーブルの単純な文字列の対訳の形で保存されているため、元の対訳文とのつながりはなくなっており、翻訳結果と元の対訳文の語を対応付けることは容易ではない。一方、EBMT では元の対訳文から動的に翻訳ルールを生成しているため、翻訳結果から元の対訳文をたどることができる。そのため、なぜシステムから出力された翻訳結果が得られたのかを説明できる。

② カスタマイズが容易であること

SMT のような翻訳モデルの学習処理が不要であるため、対訳文を追加すると、その事例はすぐに翻訳に利用されるようになる。したがって、追加した対訳文と似た文はすぐに正確に翻訳できるようになる可能性が高く、カスタマイズが容易である。

(2) EBMT のデメリット

① 翻訳速度が遅いこと

用例データベースのサイズが大きくなるにつれて、翻訳時の用例の検索にかかる時間が長くなる。現状では、同じサイズの対訳コーパスを学習した SMT よりも翻訳速度が遅い。

② 大規模化が困難であること

翻訳したい文が複雑になるにつれて、正しい翻訳を生成することが難しくなる。

1.3.3 特許分野での適用上の課題

特許文献は、一般的に文長が長く、複雑であり、また利用可能な言語資源も膨大であるため、以前の EBMT の手法で翻訳を行うことは困難であると言える。しかし、上で述べたように、入力と出力の構文木を導入した EBMT ならば、高精度に翻訳を行うことが可能である。実際 2015 年に行われた翻訳評価ワークショップ WAT2015[47]の中日特許翻訳タスクでは、京大の EBMT システムが最も精度が高かった。

また、EBMT では、対訳辞書を組み合わせることも容易である。これは文の依存構造を利用しており、用例には存在しないが辞書には収録されている専門用語が入力文に出現した

[47] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi, and Eiichiro Sumita. Proceedings of the 2nd Workshop on Asian Translation（WAT2015）, chapter Overview of the 2nd Workshop on Asian Translation, pages 1–28. Workshop on Asian Translation, 2015.

としても、その専門用語の訳を挿入するべき箇所が容易に判断できるからである。それに対しフレーズベース SMT では、このような場合に、訳語をどこに挿入するべきかを判断する材料が言語モデルしかないため、訳語としては正しくても誤った箇所に訳出される可能性が高くなってしまう。

以上の点から、EBMT は、現状では特許翻訳での実用化の実績はないものの、将来的には SMT に置き換わるポテンシャルを有していると言える。

1.3.4 多言語化の現状

従来広く使われている RBMT や、近年急速に利用が広がりつつある SMT と比べ、EBMT の実用化は遅れている。後述するように、海外での機械翻訳の研究も、近年ではほとんどが SMT に関するものであり、EBMT について多言語化の容易性を具体的に論ずることは難しい。確実なのは、EBMT では、図１．３．１-１で例示したように、対訳データの原言語と目的言語双方の文を解析して依存構造木を作成する必要があり、RBMT と同等の精度を持つ文解析技術が必要であることである。SMT においても PRE-ORDERING 技術において文解析技術や語順を並べ替える処理が必要ではあるが、語順の並べ替えの誤りは、その後の本来のフレーズベース SMT で解消される余地が残っている。その意味で、EBMT は SMT よりも高い精度の原文解析技術を必要としており、相対的に SMT よりも多言語化は難しいと考えられる。

1.3.5 翻訳精度の現状、精度向上に向けた取り組み

上述のように、2015 年に行われた翻訳評価ワークショップ WAT2015[47] における中日特許翻訳タスクでは、京大の EBMT システムが最も高い精度を示している。精度だけで見れば SMT と同等とみなすことができる。現状 EBMT に関しては、処理速度を改善し、大規模システムでの適用実績を作ることが優先的課題と思われる。

1.3.6 カスタマイズの容易性

EBMT は、SMT や RBMT と比べるとカスタマイズが容易である。ただし、単語の区切り誤りや文解析により作成される依存構造木が誤っている場合には、単に対訳データを追加するだけでは解決できない可能性があり、そのような場合には、形態素解析や構文解析の精度を改善するために、原言語や目的言語のタグ付きコーパスを作成しなければならない場合もある。

1.3.7 ドメイン適応の容易性

1.1.7 で述べたように、ここでは機械翻訳におけるドメイン適応を、汎用の翻訳エンジンとして開発されたものを、特定の分野においてより高い翻訳精度が得られるように改良する行為と考える。その場合、EBMT による特許翻訳でのドメイン適応とは、複数の技術分

野をカバーする大量の対訳コーパスがあることを前提として、さらにその上で特定の技術分野に特化した小規模の別の対訳コーパスが存在する時に、それらを組み合わせて当該分野で翻訳精度が高い EBMT を構築することになる。

SMT の場合には、対訳データを使って翻訳モデルを作成するので、適応したいドメインの少量のコーパスがあれば、そのコーパスから得られる翻訳モデルと既存の翻訳モデルを線形補完して新たな翻訳モデルを作り出す手法が提案されている。また、対訳データから求めたフレーズペアの出現回数を調整して、新たな翻訳モデルを作るといった方法も提案されている。

EBMT の場合は、当該ドメインの少量コーパスを既存の対訳コーパスに追加することはできるが、あくまでも単純に追加するだけであって、上で述べたように翻訳モデルを線形補完する際に重みを変えて合成するといった調整は難しいと考えられ、結論としてドメイン適応は SMT に比べてやりにくいと思われる。

1.3.8　新たな言語対の追加容易性

多言語化の現状のところでも述べたとおり、EBMT では対訳コーパスが必要であることに加えて、対訳コーパスの原言語と目的言語双方の文を解析して依存構造木を作成する必要があり、そのために２つの言語での解析技術を用意しなくてはならない。SMT でも PRE-ORDERING 技術を用いた語順変更をする場合には、原語側での解析技術と語順変更知識の収集が必要であるが、語順変更知識の収集は学習によってできることが既に示されているので、実質的には原語側の解析技術だけを用意すればよい。SMT でも２つの言語で双方向の翻訳を行う場合には、結局２つの言語の解析技術が必要となるが、片方向だけの翻訳で良いのであれば、必要とする解析技術が少なくて済む SMT の方が EBMT よりも言語対の追加コストは低いと考えられる。

1.3.9　ノイズに対する頑健性

原文に誤字・脱字や文法上の誤りなどがあると、処理の最初の過程で原文を解析する RBMT では、原文の解析に失敗し文全体の構造が正しく認識できずに非常に乱れた訳文を出力することがある。既に述べてきたように EBMT では、翻訳対象となる原文と、対訳コーパスから検索された類似文の原文と訳文のそれぞれを解析し、得られた依存構造木を用いて用例の組み合わせを行うため、RBMT と同様にノイズに対する頑健性は低いと考えられる。ただし、実際のところ、EBMT のノイズに対する頑健性がどの程度のものなのかは公表されている研究報告がないため不明である。学習には使われていないオープンなテストセットを EBMT と他の方式で翻訳して翻訳結果を比較・評価することにより、EBMT のノイズに対する頑健性を評価する研究の実施が望まれる。

1.4　ハイブリッド翻訳

上述の翻訳方式を組み合わせることで、より精度の高い翻訳ができるようにしたものを、ここではハイブリッド翻訳と呼ぶことにする[48] [49]。なお、翻訳の対象となっている２つの言語とは異なる第３の言語を介して翻訳を行う方式はピボット方式と呼んで区別する。ピボット方式は、例えば、日仏翻訳を行うのに、最初に日英翻訳をし、得られた英語を英仏翻訳する方式である。ピボット方式は 1.5 で述べる。

1.4.1　投票型ハイブリッド翻訳

投票型ハイブリッド翻訳とは、図１．４．１に示すように、方式の異なる複数の翻訳方式をそれぞれ独立に翻訳して得られた複数の翻訳結果の中から、最も翻訳精度が高いと推定される訳文を選択することで、最終的な訳文を得る方式である。

1.1～1.3 で述べたように、一般には、SMT は学習に使用した対訳データと同種のドメインや表現が類似した文の翻訳においては RBMT をしのぐ翻訳精度が期待できる反面、ドメインや文の類似性が低い場合には RBMT の方が翻訳精度は高いことが知られている。逆に RBMT は原文の誤りなどのノイズに対する頑健性が低いため、未知語や文法誤りの多い文においては SMT の方が概して翻訳精度が高い。投票型ハイブリッド翻訳では、それぞれ異なる方式で翻訳した結果の中から、若しくは同じ方式であっても異なる翻訳エンジンが出力した訳文の中から最も良いと推定できる訳文を選択する。そのため、原文ごとに良い翻訳結果が選定できるので、いずれか一つの翻訳エンジンから出力される訳文のみを使用する場合と比べて、最終的にはより良い翻訳結果が得られることが期待される。

[48] 潮田明:ハイブリッド翻訳のためのフレーズアラインメント. Japio 2007 YEAR BOOK
[49] 知野哲朗,釜谷聡史:ハイブリッド機械翻訳技術による日中英音声翻訳システム. 東芝レビュー Vol.64 No.2（2009）.

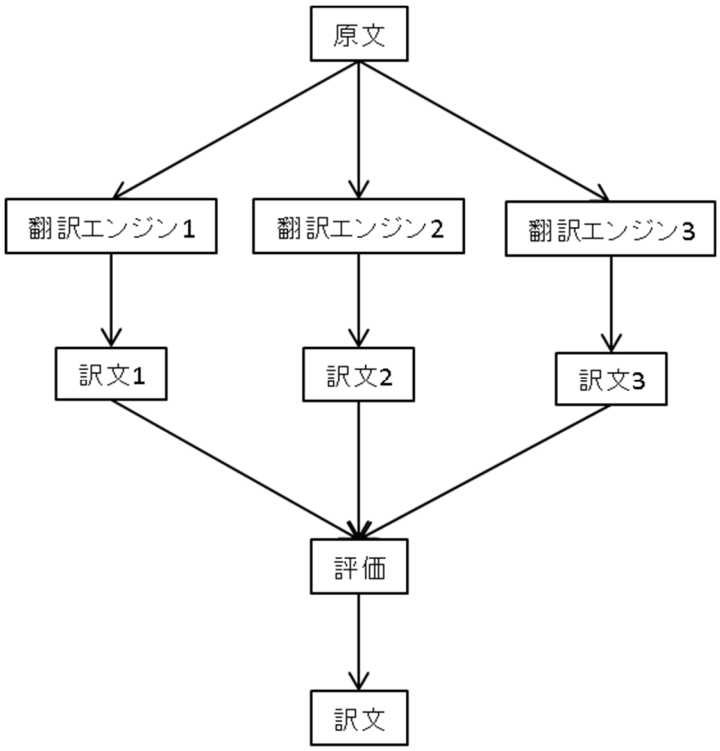


図１．４．１　投票型ハイブリッド翻訳の概要

複数の翻訳エンジンから出力された訳文の中から、最良と思われる訳文を決定する方法として最も基本的なものは、目標言語の言語モデルを使って訳文を評価する方法である。文献[50]においては、3-gram 言語モデルによって複数の翻訳エンジンからの出力を統計的な言語モデルに基づく fluency で自動評価し、最も fluency の高い訳文を選択する手法が示されている。その結果、baseline の評価と比較して 19%改善したことが報告されている。また、文献[51]においては、言語モデルと翻訳モデルのスコアを乗算した値を使うことにより言語モデル単独よりも選択精度が向上することが示されている。さらに文献[52]では順位和に基づく信頼度（RSCM）を用いてより高い精度で選択すべき訳文の判定ができるようにしている。

また、システム実装の観点で投票型ハイブリッドシステムの必要性について言及する文献もある[53]。例えば、携帯型端末に翻訳システムを実装させる際（携帯電話に音声翻訳システムを搭載させることなどがそれにあたる）に、RBMT と SMT のデータ量、処理速度、翻

50 Chris Callison-Burch and Raymond S. Flournoy: A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. In Proc. of MT Summit VIII, pp.63-66, 2001.

51 Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita: Using language and translation models to select the best among outputs from multiple mt systems. In Proc. of COLING 2002.

52 Yasuhiro Akiba,Eiichiro Sumita, Hiromi Nakaiwa, Seiichi Yamamoto, and Hiroshi G. Okuno: Using a Mixture of N-Best Lists from Multiple MT Systems in Rank-Sum-Based Confidence Measure for MT Outputs. In Proc. of COLING 2004.

53 真下修三,高京徹,趙東柱,川上健:RBMT、SMT、Hybrid MT の特徴比較と今後の展望. 言語処理学会第 21 回年次大会予稿集,pp.249-250, 2015.

訳特性等を考慮し、リアルタイム性が要求されるところではRBMTを、精度が要求されるところではSMTをそれぞれ使い分けること等が検討されている。ただし、特許文献の翻訳においてこのシステムを利用する機会があるとは考えにくい。

1.4.2　融合型ハイブリッド翻訳

融合型ハイブリッド翻訳とは、図１．４．２に示すように、一文の翻訳において、構成要素によって異なる翻訳方式で翻訳した結果を組み合わせて、最終的に一つの訳文を作成する方式である[48][49]。

最初に原文の構文解析が行われ、原文が複数（図１．４．２の例では３つ）の部分木によって構成されていることが解析される。次にそれぞれの構成要素を翻訳する。このフェーズにおいては、上で述べた投票型ハイブリッド翻訳を用いることで、それぞれの構成要素に対してベストな訳文が作成される。最後にそれらの訳をマージして、最終的な訳文を完成させる。若しくは、それぞれの構成要素に対して翻訳方式ごとに訳文を出力し、合成のフェーズでそれらの部分訳文を連結した際の最終的な文法性、自然性を評価した上で、どの部分訳文を採用するかを決定するという形を取ることもできる。

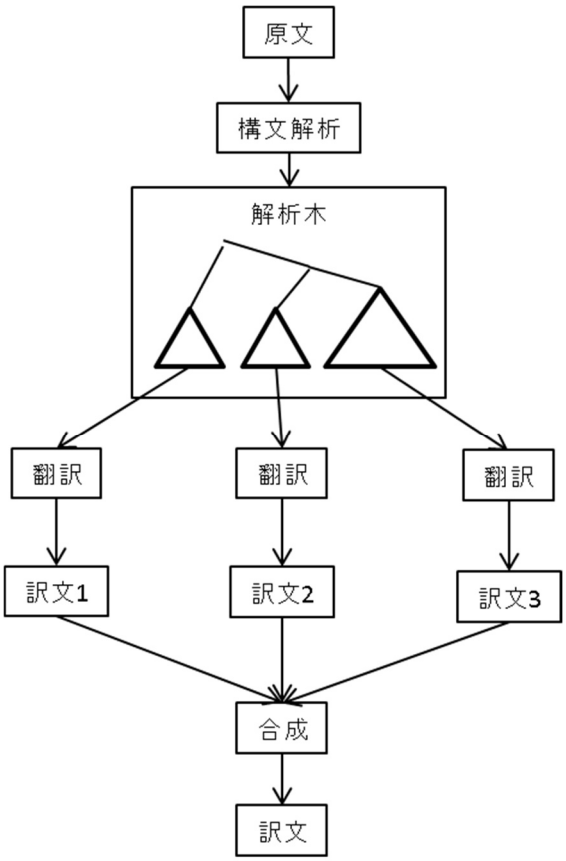


図１．４．２　融合型ハイブリッド翻訳の概要

この方式を採用するにあたり、特に RBMT と SMT を組み合わせる際に問題となるのは、RBMT におけるフレーズと SMT におけるフレーズの間に整合性が全くないことである[48]。すなわち、RBMT では構文木における構成要素をフレーズの単位とするのに対して、SMT でのフレーズは、言語学的な意味とは無関係に統計的に優位な単語の連続をフレーズとしている。これら考え方が全く違う２種類のフレーズを組み合わせて文法的に正しい単語列を出力しなければならない。文献[49]では、文法的な正確性よりも訳語や節の意味の妥当性が求められる分野（音声補翻訳等）において、融合型ハイブリッド翻訳を適用した例が紹介されている。特許分野での融合型ハイブリッド翻訳の本格的な応用はこれからの課題であると言える。

1.4.3　統計的後編集

統計的後編集とは、図１．４．３に示すように、RBMT による翻訳結果を、対訳コーパスを使って学習した結果を用いて自動修正する手法である[54] [55] [56] [57] [58]。対訳コーパスが与えられると、通常の SMT ではコーパスに収録されている原文と訳文を直接用いて、原文をどのように翻訳すべきか翻訳モデルと言語モデルを学習する。それに対し、統計的後編集では、与えられた対訳コーパスの原文を RBMT で翻訳して得られた訳文と、本来の訳文を用いて学習を行う。これにより、RBMT の出力した訳文を本来期待される訳文にするにはどのような修正を行うべきかが学習される。

---

54 江原暉将:規則方式機械翻訳と統計的後編集を組み合わせた特許文の日英機械翻訳. Japio 2010 YEAR BOOK.

55 村上仁一,徳久雅人:ルールベース翻訳と統計翻訳を結合した特許翻訳. 第 1 回特許情報シンポジウム,AAMT/Japio 特許翻訳研究会, pp. 46-53, 2010.

56 園尾聡,木下聡:統計的後編集による英日・中日・韓日特許翻訳の精度向上. Japio 2015 YEAR BOOK.

57 Terumasa Ehara: System Combination of RBMT plus SPE and Preordering plus SMT. Proceedings of the 2nd Workshop on Asian Translation（WAT2015）.

58 Satoshi Sonoh and Satoshi Kinoshita: Toshiba MT System Description for the WAT2015 Workshop. Proceedings of the 2nd Workshop on Asian Translation（WAT2015）.

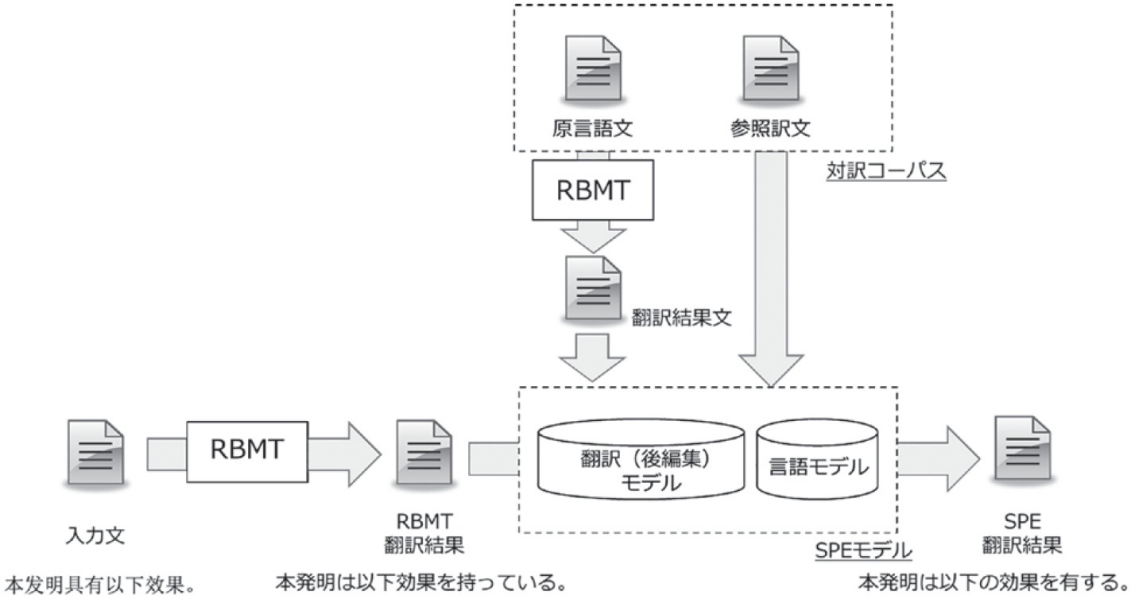


図１．４．３　統計的後編集の概要（文献 56 より引用）

この方式は、原理的には、PRE-REORDERING による SMT と等価である。すなわち PRE-REORDERING による SMT では、最初に原文の構文解析を行って語順を目標言語の順番に変更する。単語は原言語のままであるため、各単語を目標言語でどのように訳すかもすべて翻訳モデルとして学習される。それに対し、統計的後編集では、RBMT の翻訳辞書に入っている単語については目標言語の訳語に置き換えられているが、未知語についてはもとの単語のままであり、また、訳し分けの情報が不十分なものについては間違った訳語となる。しかし、これらの未知語や誤訳語も、対訳コーパスに適用可能な事例があれば本来の正しい訳語に修正される可能性があり、実験により翻訳精度が向上することが確認されている [54] [55] [56]。

この方式の特徴は、サイズが等しい対訳コーパスを用いた場合に SMT よりも高い性能が得られる点にある。異なる見方をすれば、同レベルの翻訳性能（翻訳精度、翻訳速度）を得るのに、SMT よりも少ない計算資源で済むという経済的なメリットがあるとも言える。ただし、対訳コーパスの規模が大きくなるにつれ、その差は縮まってくると考えられる。

なお、この統計的後編集という手法も、全体として見れば、RBMT や SMT に相当する翻訳方式の一種ととらえることができる。よって 1.4.1 で述べた投票型ハイブリッド翻訳における一翻訳エンジンとして、純粋な SMT エンジンと組み合わせて使用することが可能であり、単独で使用するよりもさらなる精度の向上が確認されている [57] [58] [59]。例えば、図１．４．１にあてはめると翻訳エンジン３が RBMT であった場合、訳文３の選定に関することである。

[59] 鈴木博和: 統計的後編集手法を適用したルールベース翻訳と文レベルの自動品質評価との融合. 言語処理学会第 17 回年次大会予稿集,pp.1119-1122, 2011.

1.4.4　コンセンサス型ハイブリッド翻訳

1.4.1で述べた投票型ハイブリッド翻訳は、複数の翻訳エンジンが出力した訳文の中でベストな訳文を選択する手法であった。また、1.4.2で述べた融合型ハイブリッド翻訳は、原文を解析して得られた部分構造の単位でそれぞれ異なる翻訳方式で翻訳した結果を組み合わせ方式であるのに対し、ここで述べるコンセンサス型ハイブリッド翻訳は、訳文における訳語ごとに複数の翻訳エンジンが出力したもので最適なものを選択する方式である[60] [61] [62] [63] [64]。確立された技術名称ではないが、このようにして得られる訳文を英論文ではconsensus translationと呼んでおり、それをもとに、ここではこのような名称で呼ぶことにする。

複数の翻訳エンジンの訳文を用いて一つの訳文を作り出す手法として代表的なものがconfusion networkを用いる方法である[61]。図１．４．４にその具体例を示す。図１．４．４（a）は、５つの翻訳エンジンが出力した訳文を列挙したものであり、図１．４．４（b）は図１．４．４（a）の５文の単語／フレーズ間でアラインメントを行って作成したネットワーク（又はラティス）である。ネットワークを作成する際は、まずスケルトンと呼ばれる基準となる訳文を一つ選択し、他の訳文は、それに対してアラインメントを次々と行うことにより作成される。もし対応する単語がない場合には、空を示す<epsilon>というシンボルのアークが作成される。また、各語には確率に基づく重みが付与されており、全ての訳文で一致する語には重み０が付与される。最後にネットワークの中で最もコストの小さな系列を選べば、５つの訳文の中で最もコンセンサスが得られた訳文と考えることができる。文献[61]に示されている実験結果では機械評価（BLEU）で15%スコアが改善されたと報告されている。

---

60 Srinivas Bangalore, German Bordel and Giuseppe Riccardi: COMPUTING CONSENSUS TRANSLATION FROM MULTIPLE MACHINE TRANSLATION SYSTEMS. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, 2001.

61 Evgeny Matusov, Nicola Ueffing and Hermann Ney: Computing Consensus Translation from Multiple Machine Translation Systems Using Enhanced Hypotheses Alignment. Proceedings of EACL, pp. 33-40, 2006.

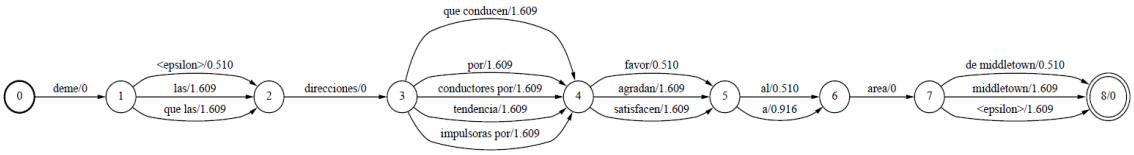
62 Wolfgang Macherey and Franz Josef Och: An Empirical Study on Computing Consensus Translations from Multiple Machine Translation Systems. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 986–995, 2007.

63 Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartz and Bonnie J. Dorr: Combining Outputs from Multiple Machine Translation Systems. Proceedings of NAACL HLT 2007, pages 228–235, 2007.

64 渡辺太郎:構文情報を直接利用した機械翻訳システムコンビネーション. 情報通信研究機構季報 58（3・4）, 63-70, 2012.

déme			direcciones	impulsoras por favor	a	área	de	middletown
déme			direcciones	por favor	a	área		
déme			direcciones	conductores por favor	al	área		middletown
déme		las	direcciones	qu e conducen satisfacen	al	área	de	middletown
déme	que	las	direcciones	tend en cia a gradan	al	área	de	middletown
*****			**********			****		**********

(a)



(b)

図１．４．４　consensus translation を作成するための confusion network
（文献 61 より引用）

Confusion network を用いる手法は文字列ベースの処理であるのに対し、文献 64 では構文的な近さを利用した手法を提案している。具体的には、複数の機械翻訳の訳文に対し構文解析を行い、それら複数の構文解析結果を表現する構文森を作成する。この構文森では、統語的な共通性が構文森における部分木の共有という形で示される。これにより、confusion network による文字列ベースの手法では扱いが難しかった、能動態や受動態の違いといった構文的な違いを考慮した処理が可能となるとしている。ただし、文献 64 における実験では、翻訳精度としては confusion network を用いた手法とほぼ同等のレベルに止まっている。

1.4.5　その他

最後に、これまで述べてきた手法とは異なる形態の SMT と RBMT の組み合わせについて述べる。

文献65では、SMT の翻訳における未知語を減らすために、RBMT の出力を用いる手法を提案している。翻訳対象である原文とそれを RBMT で翻訳した結果を対訳データとして SMT のフレーズテーブルの学習時に追加して用いることにより未知語を減らす効果があることが示されている。

65 Yu Chen, Andreas Eisele, Christian Federmann, Eva Hasler, Michael Jellinghaus and Silke Theison: Multi-Engine Machine Translation with an Open-Source Decoder for Statistical Machine Translation. Proceedings of the Second Workshop on Statistical Machine Translation, pages 193–196, 2007.

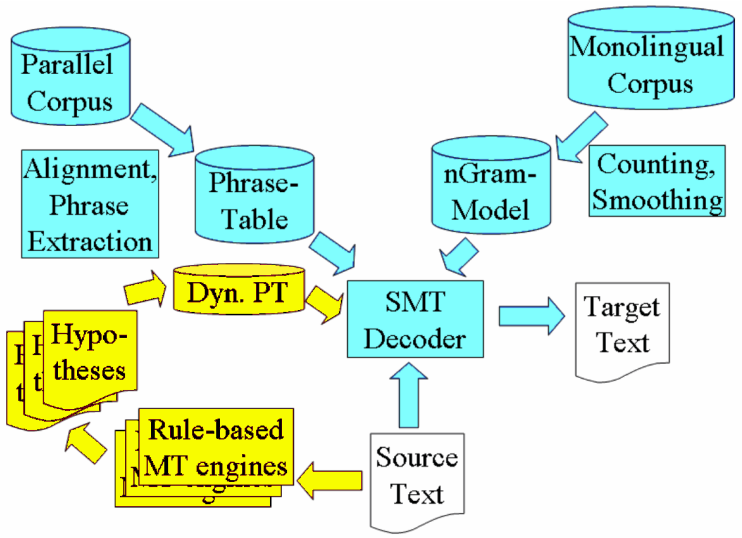


図１．４．５　RBMT の出力で補強した phrase-table を用いる SMT（文献[65]より引用）

1.5　ピボット翻訳

翻訳の対象となっている２つの言語間で直接翻訳を行える翻訳システムや２つの言語間の対訳コーパスが存在しない場合に、それらの言語とは異なる第３の言語を介して翻訳を行ったり、SMT の開発に必要な翻訳モデル（フレーズテーブル）を作成したりする方式を本報告ではピボット翻訳と呼ぶ[66]。

ピボット翻訳の手法としては、２つの翻訳エンジンをシーケンシャルに組み合わせて実現する逐次型ピボット翻訳の他に、コーパス翻訳方式、テーブル合成方式といった、主に３つの手法に分類できる[67]。逐次型ピボット翻訳は、任意の翻訳方式により実現された翻訳エンジンを実行時に組み合わせて使用する方式の最も単純な方式である。それに対し、コーパス翻訳方式は、例えば、原言語と英語の対訳コーパスがある場合、その英語を日本語に機械翻訳して原言語と日本語の対訳コーパスを作成して SMT を構築する方法である。テーブル合成方式のピボット翻訳は、例えば、原言語と英語の対訳コーパスと英語と日本語の対訳コーパスがある場合、それぞれを SMT 学習して作成した２つのフレーズテーブルから原言語と日本語のフレーズテーブルを合成して SMT を構築する方法である。つまり、後者２つの方式は必要な対訳コーパスが存在しない言語間の SMT を構築する場合に、利用可能な対訳コーパスを活用する方式である。

なお、ピボット翻訳において中間言語となる言語をピボット言語と呼ぶ。通常は、多くの言語において、英語との間の対訳辞書や対訳コーパスが入手しやすいことから、ピボット言語には英語が選択されることが多いが、最終的な目標言語が、ある言語の方言であっ

---

[66] 内山将夫,伊佐原均:統計的機械翻訳におけるピボット翻訳の比較. 言語処理学会第 13 回年次大会論文集, pp. 187-190, 2007.

[67] 三浦明波,Graham Neubig,Sakriani Sakti,戸田智基,中村哲:階層型フレーズベース翻訳におけるピボット翻訳手法の応用. 情報処理学会研究報告自然言語処理（NL） 2014-NL-219（20）, 1-7, 2014.

たり、ある言語と言語的に極めて近い関係にあったりする場合には、英語以外の言語がピボットして使われることもある。

1.5.1　逐次型ピボット翻訳

図１．５．１は逐次型ピボット翻訳により仏日翻訳を行う場合の処理内容を示した図である。この方式は、単純に第１の翻訳エンジン（この例では仏英翻訳エンジン）による翻訳結果を第２の翻訳エンジン（英日翻訳エンジン）の入力として使用する。２つのエンジンはテキスト形式の英文でつなげるだけなので、それぞれの翻訳エンジンは異なる処理方式の翻訳エンジンであっても構わない。例えば、仏英翻訳エンジンには SMT を使い、英日翻訳エンジンには RBMT を使うといったことも可能である。また、この図では、仏英、英日それぞれの翻訳エンジンが１つであるように書かれているが、それぞれは 1.4 で述べたハイブリッド翻訳により複数の翻訳エンジンを組み合わせたものであってもよい。さらにいえば、投票型ハイブリッド翻訳で用いられているような複数の訳文から、最も翻訳精度が高いと推定される訳文を選択する機構が存在するならば、第１の翻訳エンジンで複数の訳文を出力し、第２の翻訳エンジンでそれぞれの訳文を翻訳し、得られた複数の訳文から最良と思われる訳文を選択するといった処理も理論的には可能となる。

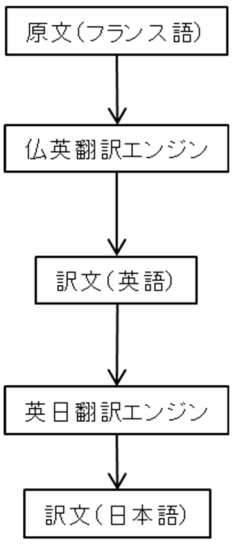


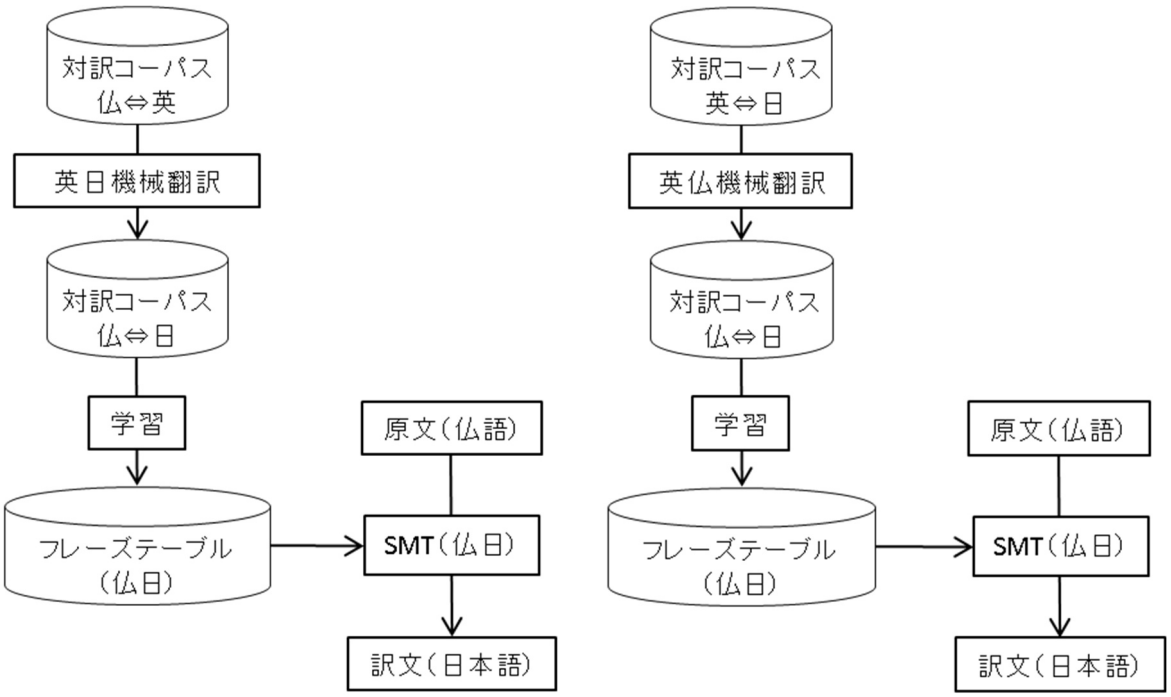
図１．５．１　逐次型ピボット翻訳

また、翻訳対象言語の一つがマイナー言語のためにピボット言語との間で十分な対訳コーパスが用意できない場合、文字ベースの SMT（Character-based SMT）を用いる手法も提案されており、有効性が示されている[68]。

[68] Jörg Tiedemann: Character-Based Pivot Translation for Under-Resourced Languages and. Domains. In EACL12, pp. 141-151, 2012.

1.5.2　コーパス翻訳方式によるピボット翻訳

図１．５．２に、コーパス翻訳方式を用いた仏日ピボット翻訳の仕組みを示す。この方式では、仏日 SMT で必要となる仏日対訳コーパスが存在しない場合に、もし仏英対訳コーパスと英日翻訳の手段があれば、それらを利用して、仏日対訳コーパスを作成し、これを使って仏日 SMT を構築する。大規模な対訳コーパス中のテキストを人手で翻訳するのは現実的ではないため、ここでの翻訳（この例では英日翻訳）も機械翻訳の利用が前提となっている。当然のことながら、機械翻訳の結果には誤りも含まれるので、これを補正する手段として、日本語として正しい大量の単言語コーパスから作成した言語モデルで翻訳結果を選定する方法がある。このように作成される仏日 SMT でも、ある程度の精度が期待できる。文献[69]では、スペイン語を中間言語としたカタルーニャ語と英語のピボット翻訳において、逐次型ピボット翻訳とコーパス翻訳方式によるピボット翻訳で有意な差はなかったとの報告がなされている。



（a）英日機械翻訳を使う場合　　（b）英仏機械翻訳を使う場合

図１．５．２　コーパス翻訳方式によるピボット翻訳

なお、図１．５．２（a）においては存在している対訳コーパスの目標言語側の言語

69 Adrià De Gispert , José B. Mariño：Catalan-English Statistical Machine Translation without Parallel Corpus: Bridging through Spanish. In Proc. of LREC 5th Workshop on Strategies for developing machine translation for minority languages,
2006.

（本例では仏英コーパスの英語）を最終的な目標言語へ翻訳して仏日対訳コーパスを作成したが、対訳コーパスの目標言語を最終的な翻訳対象での目標言語と一致させ、原言語のテキストを翻訳したい言語の原言語へと翻訳して対訳コーパスを作ることも可能である。図１．５．２（b）は対訳コーパスとして英日翻訳が存在する場合に、対訳コーパスの英語テキストをフランス語へと翻訳することで仏日対訳コーパスを作成して仏日 SMT を作っている。

以上示したように、翻訳対象とする言語間の対訳コーパスがない場合には、既存の翻訳エンジンが利用可能であれば、必要とする言語間での対訳コーパスを作ることで、SMT を開発することが可能となる。

1.5.3　テーブル合成方式によるピボット翻訳

図１．５．３に、テーブル合成方式を用いた仏日ピボット翻訳の仕組みを示す。この方式では、仏日 SMT で必要となる仏日対訳コーパスが存在しない場合に、ピボット言語である英語を一方の言語として、翻訳対象であるフランス語並びに日本語をもう一方の言語とする２つの対訳コーパスがあれば、それらを利用して仏英並びに英日のフレーズテーブルを抽出し、それら２つのフレーズテーブルを使って仏日 SMT で必要なフレーズテーブルを合成するというアプローチを取る。

文献[66]には、欧州議会の各国語での議事録から作成した Europarl 対訳コーパスを用いた実験において、本方式による翻訳精度（機械評価）が、逐次型ピボット翻訳より高いという結果が示されている。また、翻訳対象となる２言語（図１．５．３の例では仏日）間の対訳コーパスを利用した場合と比較しても、相対的な性能が 92%から 97%の水準を維持していることから、直接の対訳リソースが存在しないマイナー言語を扱う翻訳システムを実現するにあたり有望な技術であると言える。さらに文献[70]では、ピボット言語を複数用いた手法が示されている。また、文献[71]では、翻訳対象言語間の対訳コーパスを使って作成したフレーズテーブルと、ピボット言語と翻訳対象言語の間の２つの対訳コーパスから作成した２つのフレーズテーブルを、共起頻度を用いて組み合わせることにより最終的な翻訳対象言語間のフレーズテーブルを作成する手法が示されている。

---

70 Trevor Cohn and Mirella Lapata: Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora. In Proc. ACL, pp.728-735, 2007.

71 Xioaning Zhu, Zhongjun He, Hua Wu, Conghui Zhu, Haifeng Wang and Tiejun Zhao: Improving Pivot-Based Statistical Machine Translation by Pivoting the Co-occurrence Count of Phrase Pairs. In Proc. EMNLP, 2014.

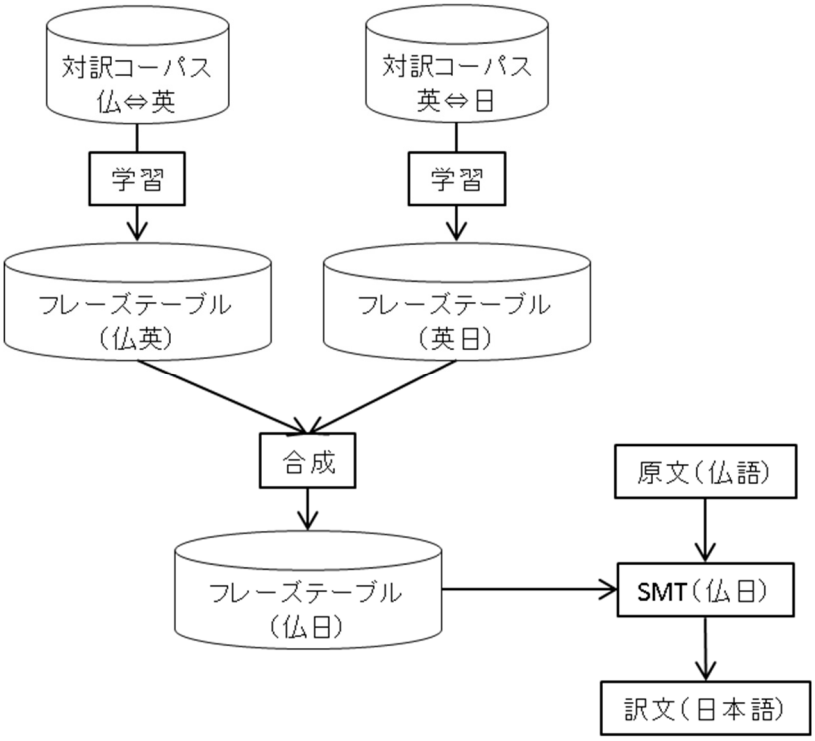


図１．５．３　テーブル合成方式によるピボット翻訳

2. 言語リソース

本節では、機械翻訳に使用することができる言語リソースの現状について述べる。なお、本節での記述は概略的なものであり、英語や日本語、中国語など言語毎の詳細な状況については「6.言語別の状況」を参照されたい。

2.1 対訳コーパス

ここでは、統計翻訳において最も必要となる言語リソースである対訳コーパスについて述べる。なお、実用システムとしての精度を考慮し、基本的にはサイズが 100 万文対以上あるものを取り上げることとする。

2.1.1 特許の対訳コーパス

特許文献を翻訳するための統計翻訳システムを開発するには、特許文献から抽出した対訳コーパスが必要である。特許自体は科学技術分野に属するため、後述する論文コーパスやマニュアルなどから作成した対訳を使うことも可能であるが、特許文献特有の言い回しもあるため、特許文献から抽出した文で作った対訳を用いた方が、翻訳精度がより高くなると予想される。

表２．１．１に英語と日本語を軸に大規模対訳コーパスの開発状況を示す。日本語と外国語間の特許コーパスに関しては、現時点では、日本国特許庁が開発した日本語と英語、中国語、韓国語間の対訳コーパスが最も大規模なものとなっている。ベトナム、タイ、インドネシア語については大規模なコーパスの存在は確認できなかった。英語と各言語との対訳コーパスについても、中国語と韓国語を除くと大規模なものは作成されていない。中英対訳コーパスは、評価型ワークショップである NTCIR-10 の PatentMT 向けに提供されたものがあり、利用目的は同ワークショップ向けの研究に限定されている。

なお、WIPO は、対訳コーパスである Corpus of Parallel Patent Applications「COPPA V2」[72]を対外提供している。学術及び民間研究機関に対し、研究目的のみの使用に限り無料でコーパスを提供しているが、研究結果を WIPO に提供することを研究機関等に求めている。商用利用を目的とする者に対しては、「再配布禁止」の条件を付した上で、2,000 スイスフランで提供している。

詳細は 6.4 中国語、6.5 韓国語、6.7 日本語を参照のこと。

[72] http://www.wipo.int/patentscope/en/data/#free

表２．１．１　特許対訳コーパスの開発状況

言語	英語との対訳コーパス	日本語との対訳コーパス
英語	-	日本国特許庁による 3.3 億文対（研究用）
中国語	NTCIR-10 の Hong Kong Institute of Education（HKIED）による 100 万文対（研究用）	日本国特許庁による 1.3 億文対（研究用）
韓国語	KIPRIS（韓国特許庁）による韓国特許の英語抄録 199 万件。（商用利用可能）	日本国特許庁による 7,000 万文対（研究用）
ベトナム語	なし	なし
タイ語	なし	なし
インドネシア語	なし	なし

2.1.2　特許以外の対訳コーパス

英日対訳コーパスとしては、国立研究開発法人科学技術振興機構（JST）と国立研究開発法人情報通信研究機構（NICT）が共同で開発した、約 300 万文対の日英論文抄録コーパス（ASPEC-JE）が最も大規模な対訳コーパスであるが、商用利用はできない。日本語と中国語、韓国語間の対訳については、100 万文対を超えるような大規模なものは、まだ作成されていない。

中英対訳コーパスについては、2,000 万文対という大規模なものが作成されている。

韓国語の対訳コーパスは科学技術分野ではないが、韓国語の新聞記事を外国語（英語、中国語、日本語）に翻訳したものをもとに作成した 100 文対規模の対訳が、言語毎に存在する。

ベトナム語、タイ語、インドネシア語については、対訳コーパスの作成が始まりつつあるが、英語との間の対訳でも規模が小さく、日本語との間の対訳コーパスは存在しない。

表２．１．２に特許以外の対訳コーパスの開発状況を示す。詳細は 6.1 ベトナム語、6.3 インドネシア語、6.4 中国語、6.5 韓国語、6.7 日本語を参照のこと。

表２．１．２　特許以外の対訳コーパスの開発状況

言語	英語との対訳コーパス	日本語との対訳コーパス
英語	-	ASPEC-JE（日英論文抄録コーパス）300 万文対（非商用）
中国語	英汉双语平行语料库（2,000 万文対。商用利用可）	なし（ASPEC-JC（日中論文抄録コーパス）68 万文対（非商用）が最大）
韓国語	新聞記事の対訳 100 万文対（商用利用可能）	新聞記事の対訳 100 万文対（商用利用可能）
ベトナム語	EVBCorpus（80 万文対。商用利用可否不明）	なし
タイ語	なし	なし
インドネシア語	OPUS プロジェクトのコーパスのうち書き言葉のもの（110 万文程度。商用利用可否不明）	なし（OPUS プロジェクトのコーパスのうち書き言葉のものは 70 万文程度。商用利用可否不明）

2.2　単言語コーパス

2.2.1　平文コーパス

構文解析・意味解析などの言語分析を加えていないテキストのみのコーパスを、平文コーパスと呼ぶ。インターネットの普及により、ロボットで Web 文書を収集することにより比較的容易に大規模なコーパスが構築できるようになったため、どの言語でも大規模なコーパスが存在する。ただし、特許文献に関して、東南アジア各国では電子化が進んでいないため、特許文献をもとにした大規模な対訳コーパスは存在していない。表２．２．１に平文コーパスの開発状況を示す。詳細は 6.1～6.7 の各国語の詳細を参照のこと。

表２．２．１　平文コーパスの開発状況

言語	状況
日本語	電子出願による大規模な特許テキストが存在する。
英語	同上
中国語	同上
韓国語	同上
ベトナム語	なし
タイ語	NECTEC word annotated corpus （500 万語のタグ付きコーパス。文数は不明。商用不可）
インドネシア語	Bahara Indonesia Newspapers Collection （900 万語のコーパス。文数は不明。商用不可）

2.2.2　注釈付きコーパス

単純なテキストだけではなく、品詞情報や構文情報など各種の言語解析の結果を注釈として付与したコーパスを、注釈付きコーパスと呼ぶ。ここでは注釈付きコーパスのうち、テキストを形態素に分割しそれぞれの品詞情報を付与した品詞タグ付きコーパスと、構文情報が付与された構文タグ付きコーパスについて述べる。これらのコーパスの作成には、言語学者や言語処理の研究者、技術者など言語処理の専門知識を持った者か、あるいは注釈作業の訓練を受けた者が行う必要があるため、平文のコーパスに比べ開発コストが高くなる。そのため、コーパスのサイズは、平文コーパスと比べると圧倒的に小さくなる傾向がある。

なお、注釈として付与される品詞・構文等は言語によって差がある。また、一つの言語でも利用の目的により付与される情報が異なるなど標準化されていないため、利用にあたっては注釈付きコーパスの仕様を検討する必要がある。

特許文献をもとにした注釈付きコーパスはほとんど存在していない。

(1) 品詞タグ付きコーパス

品詞タグ付きコーパスは、テキストを形態素に分割する際の分割位置とそれぞれの品詞についての情報が付与されたコーパスである。通常は形態素解析結果の誤りを人手で修正することにより作成される。表２．２．２‐１に品詞タグ付きコーパスの開発状況を示す。詳細は 6.1〜6.7 の各国語の詳細を参照のこと。

表２．２．２-１　品詞タグ付きコーパスの開発状況

言語	状況
日本語	京大 NAIST テキストコーパス[73]（タグ情報のみ約４万文。原文は別途入手） BCCWJ:書き言葉平衡コーパス[74]（約 17 万文。研究用途）
英語	British National Corpus（約１億語。商用利用可）
中国語	中国山西大学による Chinese POS tagged Corpus（500 万字。商用利用可能）[75]
韓国語	韓国国立国語院による韓国語情報付きコーパス（書き言葉の文が約 2,000 万文。ただし商用利用はできない。）
ベトナム語	なし
タイ語	NECTEC word annotated corpus（500 万語のタグ付きコーパス。文数は不明。商用不可）
インドネシア語	なし（インドネシア大学によるタグ付きコーパスがあるが、文数で約 2.7 万文と小規模である。）

(2) 構文タグ付きコーパス

構文タグ付きコーパスは、句構造や係り受け関係、概念依存構造などを表現するためのデータが付与されたコーパスである。品詞タグ付きコーパスと同様、構文解析や意味解析の結果得られた木構造や概念依存構造の誤りを、人手で修正することにより作成される。

英語、日本語、中国語については、比較的大規模なコーパスが作られているが、特に中国語のコーパスについては、活発に開発がなされており、種々のコーパスが提供されている。表２．２．２-２に構文タグ付きコーパスの開発状況を示す。詳細は 6.4 中国語、6.5 韓国語、6.6 英語、6.7 日本語を参照のこと。

[73] https://sites.google.com/site/naisttextcorpus/ と http://nlp.ist.i.kyoto-u.ac.jp/index.php?%E4%BA%AC%E9%83%BD%E5%A4%A7%E5%AD%A6%E3%83%86%E3%82%AD%E3%82%B9%E3%83%88%E3%82%B3%E3%83%BC%E3%83%91%E3%82%B9

[74] http://pj.ninjal.ac.jp/corpus_center/bccwj/dvd-index.html

[75] サイズが 2700 万字と、これよりも大規模な北京大学現代中国語コーパスがあるが、調査時点で商用利用の可否が不明であったため、表には山西大学のものを記載した。

表２．２．２-２　構文タグ付きコーパスの開発状況

言語	状況
日本語	京都大学テキストコーパス（約４万文）
英語	Penn Treebank（少なくとも 260 万語以上のテキストを含む。商用利用可）
中国語	Chinese Treebank 8.0（150 万語。商用利用可）
韓国語	Korean dependency Treebank（10 万文。商用利用可能）[76]
ベトナム語	なし
タイ語	なし
インドネシア語	なし

最近の動向としては、種々の言語に対して可能な限り一貫性のあるアノテーションを付与した treebank を作る目的で、米スタンフォード大などが中心となって Universal Dependencies（UD）と呼ばれるプロジェクトが立ち上がっており[77]、言語毎に定められた品詞タグセットや構文タグセットが、所定のサイトで公開されている[78]。

2.3　対訳辞書

2.3.1　対訳辞書

英日、中日、韓日翻訳のように、従来 RBMT が提供されている言語対では、100 万～300 万語程度の対訳辞書が作成され、実用に供されている。アジア言語については、英語を媒介とするピボット翻訳で日本語に翻訳する可能性もある。表２．３．１の対訳辞書の開発状況を示す。詳細は 6.1～6.7 を参照のこと。

[76] 規模は Sejong Korean Treebank の方が 15 万文と大きいが、Sejong Treebank は商用利用できないため、表には商用利用できる Korean dependency Treebank を記載した。

[77] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In Proceedings of LREC.

[78] Universal Dependencies, http://universaldependencies.org/（最終検索日:2016 年 6 月 30 日）

表２．３．１　対訳辞書の開発状況

言語	英語との対訳辞書	日本語との対訳辞書
英語	-	省略（詳細は 6.6 を参照）
中国語	省略（詳細は 6.4 を参照）	省略（詳細は 6.4 を参照）
韓国語	省略（詳細は 6.5 を参照）	省略（詳細は 6.5 を参照）
ベトナム語	なし	なし
タイ語	国際情報化協力センター（CICC）による基本語辞書（５万語。商用利用不可）	なし
インドネシア語	Rekso Translator（約 19 万語）	なし

2.3.2　パターン辞書

翻訳パターンをデータベース化したものを言語資源として公開している事例はほとんどないが、韓国語においては次のようなパターン DB が公開されており、いずれも商用利用可能である。

・名詞／名詞連語の韓英対訳パターン DB　13 万エントリー
・名詞／動詞連語の韓英対訳パターン DB　93 万エントリー
・韓英連語対訳パターン DB　50 万エントリー

2.4　単言語辞書

2.4.1　単語辞書

ルールベースの手法においては、機械翻訳の最初の段階である形態素解析において単語辞書を用いた処理が必須であったが、統計処理が主体となりつつある現在においては、形態素への分割位置や品詞付与をタグ付きコーパスから学習させることが可能であるため、単語辞書を予め用意することは必須ではなくなってきている。また、形態素解析に辞書を用いる場合であっても、形態素への分割や統語構造が正しく解析でき、語順変換が適切に行えればよいので、RBMT の場合のように大量の専門用語を訳語込みで収集し辞書に登録しておく必要はなくなってきている。したがって、大規模な単語辞書がないからといって統計翻訳のシステムを作れないということはなく、出現頻度の高い 10 万から 20 万語程度の規模の単語辞書でも実用的なシステムを開発できる可能性がある。表２．４．１に単語辞書の開発状況を示す。詳細は 6.1～6.7 の各国語の詳細を参照のこと。

表２．４．１　単語辞書の開発状況

言語	状況
日本語	UniDic（後述する形態素解析ツール MeCab 用の辞書で約 76 万語。商用利用可能）
英語	New Oxford Dictionary of English, 2nd edition（17 万語。商用利用可能）
中国語	中国語文法情報辞書（高頻度詞）（約 2.8 万語。商用利用可）
韓国語	標準国語大辞典（51 万語。商用利用については要交渉）
ベトナム語	不明
タイ語	不明
インドネシア語	KBBI（約９万語。商用利用可否不明）

2.4.2　係り受け辞書

日本語に関しては、NICT が開発した大規模な係り受けデータベース[79]が存在する。このデータベースは、約６億ページの Web 文書から抽出した 430 億文のテキストを Juman/KNP で係り受け解析した結果から、語句と語句の係り受けを抽出して、係り受けとその頻度を収録したもので、約 46 億種類の係り受けが含まれている。上で述べたように他言語においても、Treebank のような形で文全体の構造をデータ化している事例は多いが、この係り受けデータベースのような大規模なデータは見当たらない。

2.4.3　格フレーム辞書

格フレームとは、用言が持ちうる格要素の情報を用法ごとに定義したものである。一つの動詞でも自動詞と他動詞の用法がある場合には、同じ主格になる名詞でもどのような素性を持った名詞を取りうるかは用法によって異なるため、用法ごとにどのような格があるかを整理しておくことが有用である。こういった格フレームに関する情報をまとめた格フレーム辞書の一つとしては、京都大学の黒橋・河原研究室が Web テキストから自動構築した大規模な辞書がある[80]。他言語においても同様の辞書を構築する試みはなされているが、同レベルの大規模なもので公開されている事例は見当たらない。

2.4.4　N-gram データ

N-gram データは、言語モデルを作成する際の基本データであるが、単言語コーパスと形態素解析器があれば取得できるため、N-gram 単体で提供されている事例は少なく、大規模なものは Google が日本語 Web ページから抽出した N-gram データのみである。出現頻度 20

[79] Alagin HP https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html（最終検索日:2016 年 7 月 6 日）
[80] 京都大学黒橋・河原研究室 HP http://nlp.ist.i.kyoto-u.ac.jp/（最終検索日:2016 年 7 月 6 日）

回以上の 1〜7-gram を収録しており、異なり数は 1-gram で 256 万、7gram で５億 7,020 万である。言語資源協会（GSK）[81]から公開されているが商用利用はできない。

2.5　シソーラス、概念体系

シソーラスとは、単語の意味の類似関係を概念の上位・下位関係や同義・類義関係に基づいて整理、体系化した辞書である。概念体系若しくはオントロジと呼ばれるものも基本的にはシソーラスと同じであるが、概念間の関係として上位・下位や同義・類義関係以外にも、部分・全体関係といった関係等も記述対象としている場合が多い。

このような概念情報は、構文解析や概念依存関係の解析において、単語間の係り受け関係や並列関係を決定する際のエビデンスとして用いることができるため、ルールベースのシステムでは長年広く用いられてきた。しかし、近年の統計ベースの処理においては、概念的な近さや係り受け関係の強さをタグ付きコーパスから学習するため、シソーラスは必須ではなくなっている。表２．５にシソーラス、概念体系の開発状況を示す。詳細は 6.1〜6.7 の各国語の詳細を参照のこと。

表２．５　シソーラス、概念体系の開発状況

言語	状況
日本語	EDR 概念辞書（41 万語。商用利用可能）
英語	WordNet（20 万件。商用利用可能）
中国語	HowNet（６万語。商用利用可否不明。最終更新 2013 年）
韓国語	KIPRIS による技術用語シソーラスデータ（146 万件。商用利用について要交渉）
ベトナム語	Viet WordNet（５万語。商用利用不可）
タイ語	Asian WordNet（７万語。商用利用可否不明）
インドネシア語	Indonesian thesaurus Dictionary（1.6 万語。商用利用可否不明）

[81] GSK2007-C Web 日本語 N グラム第 1 版, http://www.gsk.or.jp/catalog/gsk2007-c/（最終検索日:2016 年 7 月 6 日）

3. 要素技術

本節では、機械翻訳に使用することができる要素技術の現状について述べる。なお、本節での記述は概略的なものであり、英語や日本語、中国語など言語毎の詳細な状況については６．言語別の状況を参照されたい。

3.1 形態素解析

SMT の最も基本的な方式である、句に基づく翻訳（phrase-based SMT）を行う場合でも、入力されたテキストを単語（若しくは形態素）に分解する必要がある。英語を始めとする欧州言語のように単語が空白により分かち書きされている言語では、空白によって分割すると同時に、単語に連続しているカンマ、ピリオド、クオーテーションなどを分離するだけでも統計翻訳を行うのに最低限必要な処理となるが、日本語や中国語のような膠着語の場合には、単語の切れ目を決める処理が必要になるため、統計翻訳においてもその処理を行う形態素解析（単に形態素に分割するだけの意味ではセグメンテーション）処理は必須の要素技術となる。近年では大規模なタグ付きコーパスが開発されるようになってきたことで、事前に語彙を登録した辞書を用意することなく、コーパスに出現するデータを使って形態素解析用で用いるモデルを学習して処理を行うツールも用いられるようになってきた。

表３．１に各言語における形態素解析ツールの一例を示す。形態素解析は機械翻訳に限らず、概ねすべての自然言語処理アプリケーションで必要となる技術であるため、他の技術に比べると提供されているツールの種類も比較的多い。ただし、言語によっては商用目的での利用が許可されていないものが多いため、利用にあたっては注意が必要である。

表３．１　形態素解析ツール

言語	代表的なツール
日本語	JUMAN（ルールベースの形態素解析システム） MeCab（パラメータの推定に Conditional Random Fields（CRF）を用いている。言語、辞書、コーパスに依存しない汎用的な設計になっている。商用利用可能である。）
英語	Stanford Log-linear Part-Of-Speech Tagger（商用利用可能）
中国語	Stanford Chinese Word Tagger（商用利用可能）
韓国語	HanNanum（商用利用可能）
ベトナム語	VnTokenizer（商用利用可能）
タイ語	SWATH（個人の研究者が公開しているツール。商用利用可否不明）
インドネシア語	公開されているツールはいくつかあるが商用利用であることが明示されているツールはない。

3.2　構文解析

構文解析は、RBMT において入力文の解析として必須の要素技術であるだけでなく、統計翻訳においても有用な要素技術である。句に基づく統計翻訳においては、入力文を形態素解析した結果をそのまま用いるのではなく、目標言語での語順に近づける形で原文の語順を並び替えてから訓練及び翻訳することで精度が向上することが報告されている[82]。このような語順の並べ替えは PRE-ORDERING 若しくは PRE-REORDERING と呼ばれている。例えば、英日翻訳の場合、SVO 型の言語である英語の動詞を日本語に合うように目的語の後ろに移動させる。そのような処理を行うためには構文構造の認識が必要となる。

構文解析の手法としては、従来は所定の文法規則を適用して解析を行うルールベースの手法が一般的であったが、近年では、大規模な構文タグ付きコーパスを用いて、構文解析を統計的に行う手法へと移行してきている。

表３．２　構文解析技術の開発状況

言語	代表的なツール
日本語	CaboCha（Support Vector Machines に基づく係り受け解析器。商用利用可能。ただし付属のモデルは研究利用に限定されているため、商用目的で使うためには独自にモデルを作る必要がある。） KNP（ルールベースの係り受け解析器。商用利用可能）
英語	Berkley Parser（商用利用可能） Stanford Parser（商用利用可能）
中国語	Stanford Chinese parser（商用利用可能）
韓国語	KKMA（商用目的での利用は別途協議が必要）
ベトナム語	Vitk（商用利用可能）
タイ語	特になし
インドネシア語	特になし

3.3　意味解析

意味解析とは、簡単に言えば、文の構文解析において曖昧性が存在する時に、語の意味的な情報を用いて曖昧性を解消する処理である。例えば、以下の２つの日本語文を見てみることにする。

例１：私は先月ボストンとニューヨークに行った。
例２：私は先月花子とニューヨークに行った。

[82] Terumasa Ehara: System Combination of RBMT plus SPE and Preordering plus SMT. Proceedings of the 2nd Workshop on Asian Translation（WAT2015）.

構文的に見ると、「ボストン」と「ニューヨーク」は地名、「花子」は人名であるから、助詞「と」は例１では並列助詞、例２では動作の相手を示す格助詞である。すなわち、「ボストン」と「ニューヨーク」はどちらも行き先になりうるため、これらは並列句として解釈できるが、「花子」が行先になることはないので、「花子と」の助詞「と」は、格助詞として解釈する方が妥当である。

処理を進め、文の意味を表現する形式に変換するのも意味解析の役割である。文の意味内容を表現する形式については、格フレーム若しくは述語項構造と呼ばれる形式が広く用いられている。前述の「行く」という事象の場合、動詞「行く」は動作主を持つ主格と行先を示す場所格を持つと考え、意味処理では事象を表すのに必須の要素を同定する。例えば、以下の例を見てみることにする。

例３：私は本を買った。
例４：本は買ってきたが新聞を買うのは忘れた。

係助詞「は」が使われている上の２つの例文は、構文的には同じ構造を持つが、意味解析の結果、例３の「私」は動作主、例４の「本」は目的格として解釈される。

このような意味解析も、従来は各単語に人手で付与した意味情報を用いて行われていたが、近年はコーパスを用いて、自動若しくは半自動で行われるようになっている。ただし、意味解析は、意味構造を媒介して変換を行う意味トランスファ方式の RBMT では必須であるが、SMT や EBMT では必須の処理ではない

表３．３　意味解析技術の開発状況

言語	代表的なツール
日本語	KNP（ルールベースの係り受け解析器だが、述語項解析（格解析）が可能。商用利用可能）
英語	特になし
中国語	NiuParser（商用目的での利用は別途協議が必要）
韓国語	特になし
ベトナム語	特になし
タイ語	特になし
インドネシア語	特になし

3.4　アラインメントツール

3.4.1　文アラインメントツール

文アラインメントツールとは、特許文献のパテントファミリーのように、文単位では対応が取れていないが、文献単位では対応することが分かっている２つの文書から、対応す

る文を見つけ出して対訳コーパスを作成するためのツールである[83]。

・hunalign[84]

ハンガリーの Media Research and Education Center を中心に開発されたツール。文の長さの情報を利用して文対応を推定する。対訳辞書があれば、その情報を利用することもできる。文の対応付けにおいては、文の出現順序は同じであるという前提に基づいて処理を行っているため、２つの言語の特許文書で、言語 A での請求項は文書の先頭、言語 B は文書の最後にあるような場合には、それらの対応付けを行うことはできない。GNU LGPL v2.1 でのライセンスにより、商用利用可能である。

・Microsoft Bilingual Sentence Aligner[85]

米マイクロソフトが開発したツール。文の長さと単語対応の２つを併用して文対応を推定する。2003 年以降は更新されておらず、商用利用もできない。

3.4.2 単語アラインメントツール

単語アラインメントツールとは、文同士の対応が取れている対訳コーパスを入力とし、各対訳文の中で対応する単語若しくは句を見つけ出すツールである。句に基づく翻訳（phrase-based translation）では、このツールを用いて対応する句を見つけ出し、phrase-table と呼ばれる翻訳モデルを作成する。

・GIZA++[86]

ドイツのアーヘン工科大学で開発された単語アラインメントツール。IBM モデルに加え HMM（隠れマルコフアラインメント）モデルが実装されている。また、種々の smoothing 手法も実装されている。GPL でライセンスされ商用利用可能である。

・mgiza[87]

CMU で開発された本ツールは GIZA++に代わるツールとして開発されたもので、アラインメント処理をマルチスレッドにより並列化することで処理速度の高速化を図っている。

---

83 Omar Zaidan and Vishal Chowdhary. 2013. Evaluating（and Improving） Sentence Alignment under Noisy Conditions. Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 484–493.

84 The hunalign sentence aligner, https://github.com/danielvarga/hunalign（最終検索日:2016 年 6 月 30 日）

85 Bilingual Sentence Aligner, https://www.microsoft.com/en-us/download/details.aspx?id=52608（最終検索日:2016 年 6 月 30 日）

86 GIZA++: Training of statistical translation models, http://www.fjoch.com/giza-training-of-statistical-translation-models.html（最終検索日:2016 年 6 月 30 日）

87 GitHub - moses-smt/mgiza: A word alignment tool based on famous GIZA++, extended to support multi-threading, resume training and incremental training. https://github.com/moses-smt/mgiza （最終検索日:2016 年 6 月 30 日）

・Berkeley Aligner[88]

対訳コーパスから単語対応を推定するツールである。教師ありと教師なしの２種類の推定モードで動作させることができる。学術目的のみでの利用が可能である。

3.5　言語モデルツール

言語モデルは、特定の言語においてある文字列、単語列あるいは文などの表現が与えられたときに、その表現がその言語においてどれくらい自然な表現であるかを判定するために用いられる。統計翻訳においては、デコーダが生成する数多くの翻訳候補の中から、目標言語として最も確からしい候補を選択するのに用いられる。代表的な言語モデルとしてN-gram モデルが広く用いられているが、他にも隠れマルコフモデルや最大エントロピーモデル、最近ではリカレントニューラルネットワーク（RNN）モデルなども用いられている。

・KenLM Language Model Toolkit[89]

エジンバラ大学の Kenneth Heafield らによって開発されたツールキットで、統計翻訳開発用のパッケージである Moses に同梱されているのを始め、広く使われているツールである。N-gram モデルに基づく言語モデルに対し、頑健な確率付与を行うための手法としてKneser-Ney 法による平滑化（smoothing）処理が含まれている。以下に述べる SRILM と比較し、少ないメモリ量で高速に動作する。LGPL ライセンスでの商用利用可能である。

・SRILM[90]

SRI インターナショナルが開発した言語モデルのツールキットである。N-gram モデルに基づく言語モデルに対し、平滑化手法として Good-Turing 法、Witten-Bell 法、Kneser-Ney 法などの平滑化処理が実装されている。基本的には学術目的での利用が認められているが、商用利用向けのライセンスもある。

・IRST Language Modelling（IRSTLM） Toolkit[91]

イタリアの研究機関 Fondazione Bruno Kessler によって提供されている言語モデルツール。Moses にも同梱されている。LGPL v2 でのライセンスにより商用利用可能である。

88 Berkeleyaligner, https://code.google.com/archive/p/berkeleyaligner/（最終検索日:2016 年 6 月 30 日）
89 KenLM Language Model Toolkit, http://kheafield.com/code/kenlm/（最終検索日:2016 年 6 月 30 日）
90 SRILM - The SRI Language Modeling Toolkit, http://www.speech.sri.com/projects/srilm/（最終検索日:2016 年 6 月 30 日）
91 IRSTLM, http://hlt-mt.fbk.eu/technologies/irstlm（最終検索日:2016 年 6 月 30 日）

・RandLM[92]

分散ハッシュテーブルを用いることで、複数のマシンに言語モデルを分散配置することができ、大規模な言語モデルを扱うことが可能である。GNU GPL v3 でのライセンスにより商用利用可能である。

・Berkeley LM[93]

米カリフォルニア大学バークレー校で開発された言語モデルツール。上で述べた KenLM と同等の処理速度を持つ。Apache License 2.0 でのライセンスにより商用利用可能である。

・RNNLM toolkit[94]

ブルノ工科大学で開発されたリカレントニューラルネットワーク（RNN）を使った言語モデルのツールである。評価型ワークショップのテストセットを用いた評価により N-gram ベースの言語モデルよりも機械評価で高いスコアが得られることが確認されている。

3.6 統計的機械翻訳システム

3.6.1 Moses[95]

研究目的のみならず商用ツールとしても広く使われている SMT システムである。上述の GIZA++、KenLM, SRILM 等のツールをシステムに組み込んでおり、句に基づく翻訳（phrase-based translation）、階層型の句に基づく翻訳（hierarchical phrase-based translation）、構文に基づく翻訳（syntax-based translation）の３種類の方式の SMT を行うことができる。言語モデルツールなど第三者が開発したツールも用いられているが、対訳コーパスを用意すれば、このツールを用いて SMT で必要となる言語モデルと翻訳モデルの作成、パラメータの調整を行うことができ、できたモデルとパラメータを用いて翻訳を行うことが可能である。ツールの最新更新は 2015 年 12 月である。LGPL でライセンスされており商用利用可能である。

3.6.2 情報通信研究機構による統計翻訳システム[96]

国立研究開発法人情報通信研究機構（NICT）が開発した統計翻訳システムが、同機構が自ら提供している翻訳サービス「みんなの自動翻訳@TexTra®」にて利用されている。有償で商用利用可能である。

---

92 RandLM, https://sourceforge.net/projects/randlm/（最終検索日:2016 年 6 月 30 日）
93 Berkeleylm, https://code.google.com/archive/p/berkeleylm/ （最終検索日:2016 年 6 月 30 日）
94 RNNLM Toolkit, http://rnnlm.org/（最終検索日:2016 年 6 月 30 日）
95 Moses, http://www.statmt.org/moses/（最終検索日:2016 年 6 月 30 日）
96 みんなの自動翻訳@TexTra®, https://mt-auto-minhon-mlt.ucri.jgn-x.jp/（最終検索日:2016 年 6 月 30 日）

3.6.3　株式会社 NTT データ及び日本電信電話株式会社による翻訳サービス[97]

日本電信電話株式会社が開発し、株式会社 NTT データが提供する統計的機械翻訳技術を利用した翻訳サービスである。有償で商用利用可能である。

3.6.4　Travatar[98] [99]

NAIST の Graham Neubig 氏らによって開発されたシステムであり、構文木に基づく SMT が可能である。さらに圧縮統語森の利用もできる。翻訳モデルの学習、言語モデルの学習、パラメータのチューニング、翻訳処理からなる一連の作業が可能である。ただし、単語アライメントは GIZA++など他のツールを利用して行う。LGPL で利用可能である。

[97] 技術文書を対象にした法人向け機械翻訳サービスを提供開始,
http://www.nttdata.com/jp/ja/news/release/2015/012702.html（最終検索日:2016 年 6 月 30 日）

[98] Graham Neubig : Travatar: A Forest-to-String Machine Translation Engine based on Tree Transducers, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 91-96, 2013.

[99] Travatar, http://www.phontron.com/travatar/index.html（最終検索日:2016 年 6 月 30 日）

4. 精度評価

4.1 人手評価

4.1.1 評価方法の概要と特徴

機械翻訳の結果を人間が評価する基準として、長年、表４．１．１に示す adequacy（正確性）と fluency（流暢性）という２つの基準が広く使用されてきたが、近年、特許文献の機械翻訳を評価する基準として新たな基準が提案されている。

表４．１．１　adequacy と fluency の評価値[100]

	adequacy
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

	fluency
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible

一つは、評価型ワークショップである NTCIR における機械翻訳の評価タスクで用いられている acceptability であり[101]、もう一つは特許庁による特許文献機械翻訳の品質評価手順である[102] [103]。

NTCIR で提案された acceptability は、従来の adequacy の代わりに用いられており、adequacy が原文の内容が訳文にどの程度反映できているかを５段階で評価するのに対し、acceptability は図４．１．１に示す評価手順のフローチャートから分かるように、原文の意味内容が正確に訳出されているかと同時に、訳文の文法の正確性や自然性も評価できるようになっている。

100 Phlipp Koehn: Statistical Machine Translation p219

101 Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou: Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, Proceedings of NTCIR-9 Workshop Meeting, Tokyo, Japan, pp.559–578, 2011.

102 特許庁:特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査、平成 23 年 2 月。

103 特許庁:特許文献機械翻訳の品質評価手法に関する調査、平成 26 年 2 月。

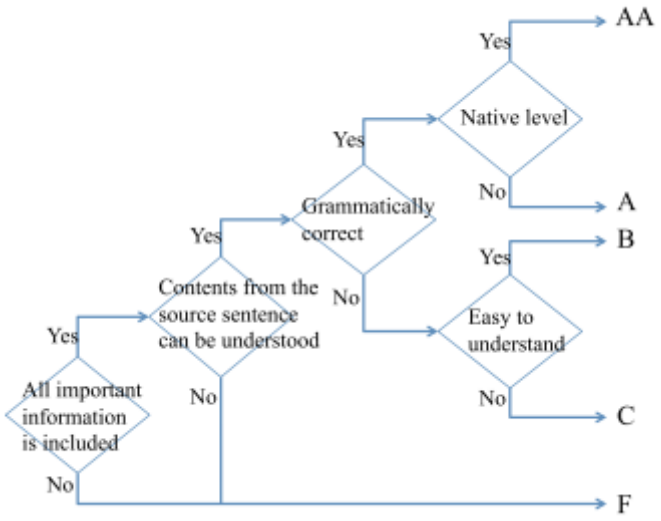


図４．１．１　acceptability の評価手順（文献 [101] より引用）

一方、特許庁による品質評価手順は、「内容の伝達レベルの評価」と「重要技術用語の翻訳精度の評価」の２つの評価基準によって構成されている。難解な技術用語を含み、長文が多い特許文献の翻訳においては、現状の機械翻訳では十分な精度で翻訳できない事例が多いため、翻訳結果に誤りが含まれている場合でも、文献を理解する上でどの程度役に立ちうるかを、従来の基準よりも正確に評価したいという意図が反映された基準になっていると言える。

なお、上で述べた評価手法は、人間が翻訳した翻訳結果を評価する際にも用いることができるが、機械翻訳というコンピュータによる処理が適切に行われているかの観点で翻訳結果を評価する仕方もある。例えば、過去に JEIDA（日本電子工業振興協会）が提案した機械翻訳システムの評価基準[104]には、機械翻訳システムの開発者がシステムの性能を評価する際にチェックすべき技術的項目を定め、それに基づいて評価を行う手法や、翻訳対象となっている言語間の翻訳において、解決すべき言語現象の視点で評価文を人工的に作成し、訳文からその言語現象の適切な処理が行われているかを読み取って評価する手法が提案されたこともある。しかし、前者の場合はシステムの内部処理に精通している人間しか評価できないこと、後者の場合には個々の言語現象に対する処理の網羅率と MT を実文に適用した場合の精度や有用性との関係が分からないと、ユーザが評価したいと考えているドメインでどの程度の精度が期待できるかを評価することが難しいことから、上で述べたような翻訳結果だけを用いて評価する手法のように広く普及するには至っていない。

[104] 井佐原均,新納浩幸,山端潔,森口稔,野村浩郷:JEIDA 機械翻訳システム評価基準（品質評価編）-英日翻訳の品質評価項目の検討と評価用コーパスの作成-．情報処理学会研究報告自然言語処理（NL） 1993（61（1993-NL-096））, 81-88, 1993-07-09

4.1.2　人手評価の課題

人手評価の代表的な課題には、少なくとも次の３つがあると考えられる。

・評価コスト
・評価の揺れ
・評価文の選定

ただし、評価文の選定に関する課題は、自動評価でも同じであるので、4.3 にまとめて記載する。

(1) 評価コスト

後で述べる自動評価（機械評価）に比べると、人間による評価作業は正確性が高いことが期待できる反面、機械での評価に比べ評価作業に時間がかかる。そのため、評価文の数を増やすためには作業時間の確保が必要であり、コストに直結する。単一のシステムの絶対的な評価であれ、複数のシステムの相対的な評価であれ、信頼性の高い評価を行うためには、ある程度の量の評価は必要であるため、そのための予算を確保しなければならないことになる。

このような評価コストの問題を軽減するため、以下のような工夫がなされている。一つは、レファレンス訳文だけを見て評価する方法、もう一つはクラウドソーシングによる評価の実施である。

レファレンス訳文だけを見て評価する方法とは、文字通り、機械翻訳の訳文を評価するのに参照訳だけを見て評価する方法である。通常の adequacy 評価に類する評価では、最初に原文を読んでその内容を理解し、その上で翻訳結果を読んで、原文での意味内容が適切に訳出されているかを評価する。そのため、原文を正しく理解する言語能力が必要である。例えば、多くの日本人は、中学以降の教育において長年英語を勉強するので、英日翻訳の評価ができる技術者や研究者を確保することはそれほど難しい問題ではない。しかし、それと同じことを、中日翻訳や韓日翻訳で行おうとしても、中国語や韓国語ができる技術者や研究者は極めて少なく、広い技術分野で大規模に行おうと思っても現実には不可能である。特許庁の提案する評価基準において、「内容の伝達レベルの評価」が、人手翻訳（参照訳）の内容に照らして評価するようにしているのは、この問題に対する一つの解決策といえる。

一方、クラウドソーシングによる評価とは、インターネット経由で労働力を通常よりも安価に提供する多数の作業者（これをクラウドワーカーと呼ぶ）に評価作業を行わせる手法である。専門性の高い翻訳が行える翻訳者に評価を依頼するのと比べ、より安価な価格で作業をしてもらうことが期待できる。しかし、その反面、より多くの作業者が評価作業にかかわることになるため、下で述べる評価の揺れの問題に加え、それぞれの評価者の専

門性を含めた作業品質の確保が課題となる。このような問題に対処するために、評価課題それぞれに対して評価者を 3 名以上用意し、多数決により最終的な評価結果を決定するといった方法も用いられている[105] [106]。

なお、クラウドソーシングの利用法として、人手評価対象のシステム数が多い場合に、クラウドソーシングによって一次フィルタリングを行って上位 N 件に対象を絞り込み、その後に、よりコストをかけた人手評価を行うという形で利用されることもある[105] [106]。

(2) 評価の揺れ

自動評価と比べ、人間による評価作業においては、全く同じ評価課題に対して評価結果が異なる場合が出てくる。これが評価の揺れの問題である。

評価の揺れには大きく分けて次の 2 種類の揺れがある。

・評価者が異なることによる評価の揺れ
・同一評価者が異なる判断をすることによる評価の揺れ

一つは、評価者による揺れである。例えば、特許庁による「内容の伝達レベルの評価」では、表４．１．２のように、原文で記載されている重要情報が訳文において何%訳出されているかで判断する。原文における重要情報と重要でない情報の区別、訳出されている割合はすべて評価者の主観によって決まるため、事前に事例などで目安を示したとしても、実際の評価では評価者によって異なる評価をつける事例が生じる。そのような評価の割れの際に、評価者同士で評価が同一になるように調整することも評価手順としては可能だが、その調整作業も評価コストに含まれることになるため、常にそういった調整ができるわけではない。

表４．１．２　内容の伝達レベルの評価基準[107]

5：すべての重要情報が正確に伝達されている。（100%）
4：ほとんどの重要情報は正確に伝達されている。（80%～）
3：半分以上の重要情報は正確に伝達されている。（50%～）
2：いくつかの重要情報は正確に伝達されている。（20%～）
1：文意がわからない、もしくは正確に伝達されている重要情報はほとんどない。（～20%）

[105] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi and Eiichiro Sumita: Overview of the 1st Workshop on Asian Translation, 2014.
[106] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi and Eiichiro Sumita: Overview of the 2nd Workshop on Asian Translation, 2015.
[107] 特許庁 特許文献機械翻訳の品質評価手順書

もう一つは、同一評価者であっても、評価対象訳文を見る順序や作業による疲れなどにより、同じ訳文に対しても異なる評価をしてしまうケースである。このような揺れの問題を軽減するためには、複数の翻訳システムを評価する場合に訳文の提示順を適当に変更するといった工夫が必要である。また、同一の課題をある間隔を置いて評価させ、評価結果に差が生じていないかをチェックするような工夫も必要となる。

いずれにしろ、人手評価を行う際にはこのような評価の揺れは必ず発生するため、評価結果をまとめる段階で、統計処理により信頼区間を求めて使用することが好ましい[105][106]。

4.2 自動評価

4.2.1 自動評価手法の概要と特徴

人手評価における高コストの問題や評価の揺れの問題を解決するために提案されたのが、コンピュータにより機械的に評価を行う自動評価（若しくは機械評価）の手法である[108][109][110][111][112]。以下では、代表的な５つの手法を紹介するが、これらの手法では、機械翻訳が出力した訳文と比較するための参照訳を人間による理想訳として用意しておき、それらとの一致度を測ることで翻訳精度を評価している。

(1) BLEU

自動評価手法として初期に提案され、現在でも広く用いられている評価法である[113]。参照訳と翻訳システムの結果の間で一致する単語の N-gram の数に基づいてスコアを算出する。N は１から４までのものを使用する場合が多い。実装が簡単で処理速度も速いが、せいぜい 4-gram という短い単語の連続にしか着目しないため、日本語と英語間のように語順が大きく入れ替わるような言語間での翻訳結果を評価する場合で、しかも目標言語が日本語のように語順の自由度が大きい言語の場合には後述の自動評価の課題で述べるように人間の評価との間の相関が低くなる傾向がある。特に「鉛非含有」と「鉛含有」のように意味的に正反対でも文字列の上では似ている場合がある。そのため、自動評価値が良い場合でも、文の意味は完全に誤っている場合がある。評価値は 0.00（悪い）～1.00（良い）値をとる。

---

108 江原他: 機械翻訳精度の各種自動評価の比較. Japio 2009 YEAR BOOK.

109 Graham Neubig:文レベルの機械翻訳評価尺度に関する調査. 情報処理学会第 212 回自然言語処理研究会, NL-212, 2013

110 越前谷博,磯崎秀樹:自動評価法を用いた機械翻訳の定量的評価. 第 3 回特許情報シンポジウム, 2014.

111 平成 22 年度 特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査報告書,4.1 節

112 平成 27 年度特許審査関連情報の日英機械翻訳文の品質評価に関する調査報告書,5.2 節

113 Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu : BLEU : a Method for Automatic Evaluation of Machine Translation, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, 2002.

(2) NIST[114]

NIST（National Institute of Standards and Technology）は、BLEU とよく似た自動評価手法であり、機械翻訳文の精度を適合率の点から評価する。NIST の値は 0（悪い）以上で良い方の値の上限の値はなく、大きければ大きいほど機械翻訳の精度が高いことを示す。

$$NIST = BP_{NIST} \sum_{n=1}^{N} q_n$$

ここで qn は

$$q_n = \frac{\sum_{w_1,w_2,\ldots,w_n \in G_n} \mathrm{Info}(w_1, w_2, \ldots, w_n)}{\sum_{w_1,w_2,\ldots,w_n \in G_n} 1}$$

図４．２．１　NIST の算出式

平成 27 年度特許審査関連情報の日英機械翻訳品質評価に関する調査より[115]

(3) TER

TER（Translation Edit Rate）は、システムの翻訳結果を参照訳に一致させるために行わなければならない編集作業の量によって翻訳結果を評価するものである[116]。ここでいう編集作業とは語の追加、削除、修正、移動などであり、機械翻訳の訳文を正確かつ自然な訳文に近づけるために行うべき後編集の作業量が少なければ少ないほど良い訳文であるという前提に基づいた評価法である。

(4) IMPACT

最長共通部分単語列（Longest Common Subsequence: LCS）に基づいて評価スコアを算出する[117]。参照訳と共通する最も長い単語列を見つけ、残りの部分に対して再帰的に処理を行うことで、人手評価との高い相関性のある評価結果が得られている。また、同じ研究者らから、より大局的な観点での評価が可能な手法も提案されている[118]。BLEU における語順と文レベルの構造に関する問題を解決するために提案された。

114 NIST : Automatic Evaluation of Machine Translation Quality Using N-gram Co-OccurrenceStatistics, http://www.nist.gov/speech/tests/mt/doc/ngram-study.pdf, 2002

115 https://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kikai_honyaku/h27_03.pdf　p88

116 Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J.: A study of translation edit rate with targeted human annotation, pp. 223–231（2006）.

117 Hiroshi Echizen-ya and Kenji Araki: Automatic Evaluation of Machine Translation based on Recursive. Acquisition of an Intuitive Common Parts Continuum. In Proceedings of the MT Summit XII Workshop on Patent Translation, pp. 151–158. 2007.

118 越前谷博,荒木健治:機械翻訳のための自動評価法における文分割を用いた大局的評価の利用．言語処理学会 第 22 回年次大会 発表論文集, pp.609-612, 2016.

(5) RIBES

RIBES（Rank-based Intuitive Bilingual Evaluation Score）は、参照訳とシステムの翻訳結果との間で共通して出現する単語の出現順序に着目した評価手法である[119] [120]。この手法も上で述べた IMPACT 同様、BLEU における N-gram ベースの局所的な評価の問題点を解決するために提案されたもので、日本語と英語のように語順が大きく異なる言語対に対しても人間による評価結果と高い相関を持つことが日英特許翻訳に対する実験で明らかにされている[120]。

4.2.2 自動評価の課題

上述の代表的な自動評価手法において触れたとおり、自動評価を人手評価の代わりとして使うという視点で考えた場合、人手評価の結果と自動評価の結果の間に高い相関があることが望ましい。すなわち、例えば、３つの翻訳システムを自動評価と人手評価で精度を順位付けした場合に、それらの順序が一致するのが理想である。しかし、これまで発表されている研究成果を見る限り、人手評価と自動評価の間に常に正の相関が認められるという状況には至っておらず、報告によっては負の相関があると示されている事例もある[106] [121]。

これまで述べたように、従来の知見では、語順に関しての制約が少ない日本語と、文法による語順の制限が強い欧米言語の間の機械翻訳に対する評価においては、RIBES のように並べ替えの情報を重視する手法が必要であると思われてきた。しかし、文献[121]では韓国語と日本語のように語順がほとんど同じ場合でも機械評価と人手評価の相関性がない事例が示されており、自動評価を人間による評価の代わりに利用するためには、さらなる研究の進展が望まれる。

特許文献の機械翻訳文の自動評価に使用する参照訳の作成方法として、特許文献のパテントファミリーから作成された対訳コーパス利用する方法がある。パテントファミリーの対訳は、意訳されていたり、各国の特許文献の定型にしたがった言い回しで書かれていて対訳としては正くない場合があったりすることため、参照訳として使えるかどうか、専門家が確認する必要である。

4.3 評価セットの選択

特許文献を機械翻訳する場合、翻訳者の参考用資料として使用するのか、文献の概要把握するために使用するのかなど、機械翻訳システムの利用目的に応じて、評価に使うべき原文の種類が異なる。したがって、人手評価であれ自動評価であれ、評価結果の信頼性を

---

[119] 平尾努,磯崎秀樹,Duh Kevin,須藤克仁,塚田元,永田昌明：RIBES:語順相関に基づく翻訳の自動評価法, 言語処理学会 第 17 回年次大会 発表論文集, pp.1115–1118, 2011.

[120] 平尾努,磯崎秀樹,須藤克仁,Duh Kevin,塚田元,永田昌明:語順の相関に基づく機械翻訳の自動評価法,自然言語処理,Vol.21, No.3, pp.421–444, 2014.

[121] 中澤敏明,園尾聡,後藤功雄:特許文の中日・韓日機械翻訳の人手評価結果の分析. 平成 27 年度 AAMT/Japio 特許翻訳研究会報告書

高めるためには、評価に使用する文が適切に選択できていることが望ましい。

翻訳評価ための評価文のセットを決定するにあたっては、次のような点を考慮して評価文を決める必要がある。

4.3.1 パテントファミリー

新たな参照訳を作成せずに、パテントファミリーの文献から採取した対訳文をテストセットとして用いることがあるが、SMT は同じ方法で作成した対訳文を学習データとしている場合がある。評価文のセットが学習済みのデータに含まれていると正当な評価ができないので、評価セットと SMT の学習データの関係に注意する必要がある。

4.3.2 技術分野

多様な技術分野において機械翻訳サービスを提供しようと考えている場合には、ある粒度で技術分野を分類し、その分類ごとに評価セットを定めて評価し、精度がどの分野でも同じ程度であると確認できることが望ましい。4.2.2 で述べたように、自動評価と人手評価の間には必ずしも相関関係が認められない場合もあるが、技術分野を比較的細かな単位に分類した場合でも、大量の評価を短時間で実施できる自動評価のメリットを活かすことで、それらの技術分野における翻訳精度の相対的な高低はある程度把握でき、それをもとに重点的に精度改善を行うべき技術分野を特定することが可能となる。

4.3.3 文の長さ

評価セットに収録する評価文の長さの分布は、当然のことながら、翻訳対象文書全体での文の長さの分布を反映したものであることが望ましい。実際にはほとんどが 40 文字以上の文であるにも関わらず長さが 40 文字未満の文だけを集めて評価セットを作った場合には、いくら評価セットに対する評価で翻訳精度が高いと分かっても、実文を翻訳した場合には期待されるレベルの精度が得られないことになってしまうからである。そのため、評価セットに収録する文の選定にあたっては、文の長さの分布を考慮する必要があり、すでに特許庁の調査研究においても、この点を考慮した選定がなされたこともある [102]。

4.3.4 文の種類

(1) 平叙文、疑問文などの種類

通常、文の種類というと、平叙文や疑問文、命令文などが想起されるが、一般の文書と違い、特許文献に出てくる文はほとんどが平叙文であるため、これらの文の種類に関しての分布は考慮する必要はないと考えられる。

(2) タイトルのような名詞句と通常の文の違い

特許文献のタイトルは通常名詞句であり日本語に翻訳したときにも当然名詞句であるべ

きである。したがって、文の長さの問題と同様、明細書における翻訳対象テキストの数における通常の文とタイトルのような名詞句の出現比率に応じた割合でテストセットに含めることが望ましい。

(3) 特許請求の範囲とそれ以外の文

論文や製品マニュアル等の一般の技術文書と比較して、特許文献における請求項の記載は、それ以外の文と長さや複雑性の点で大きく異なっている。そのため、上で述べたタイトルと同様、実際の明細書における請求項の文数の比率に従って、評価セットの文に請求項から抽出した文を含めることが望ましい。もちろん、4.3 の冒頭で述べたように、請求項を含めるべきか否かは、機械翻訳の利用目的によって異なる。すなわち、要約だけを翻訳対象とする翻訳システムにおいては、請求項は翻訳する必要がないため、このような場合には評価セットに加える必要がないためである。

(4) 文の評価と文献の評価

特許文献の翻訳では一つの文献の中で採用される訳語が統一的であることが望ましいがここまで説明してきた評価は訳語の統一を評価の対象としない。また、将来の技術になるが、例えば、日英翻訳の場合、文脈解析によって原文で省略された主語を補完すること、照応関係を推定すること、そして定冠詞・不定冠詞を使い分けて翻訳することが望まれる。しかし、これらの評価は１文単位では難しいので、評価の単位をパラグラフ単位や１文献単位に広げることが考えられる。

このように評価の単位を１文単位からパラグラフや１文献単位に変えることにより、１文単位の評価では僅差である日英方向の SMT と RBMT の優劣が変化する可能性がある。例えば、１文献単位の評価では訳語のコントロールが容易な RBMT の方が高い評価を受ける可能性がある。

4.4　絶対評価と相対評価

4.4.1　絶対評価

絶対評価とは、評価対象となっている複数の機械翻訳システムに対し、個々に絶対的な評価スコアを付与することにより行われる。例えば、4.1 の人手評価のところで触れた adequacy や fluency による評価尺度や特許庁による「内容の伝達レベルの評価」の評価基準を用いて評価すれば、その評価尺度での平均スコアを絶対評価の値として得ることができる。

このような評価を行えば何らかのスコアは得られるが、このスコアだけで新たな技術分野や言語に対する機械翻訳システムの導入可否が判断できるわけではない。対象となる言語に対する既存の機械翻訳システムの有無や、そもそも対象言語における技術翻訳者の数などにより、機械翻訳に期待される精度は異なってくるためである。そのため、機械翻訳

サービスの多言語化を進めるには、評価法だけでなく、そのような言語の特性や機械翻訳に期待される精度なども事前に調べておく必要がある。

4.4.2 相対評価

相対評価とは、複数の機械翻訳システムの間で翻訳性能、より狭義には翻訳精度によって順位付けを行うことを指す。仮に翻訳精度が必ずしもすべてのユーザを満足させうるレベルでない場合でも評価の時点で最良のシステムを決定しなければならないような場合に必要となる。

単純には、すべての対象システムを絶対評価し、その評価結果に基づいて順位を付ければ相対評価となる。しかし、人手評価で adequacy や fluency を求めるには決められた基準できちんとした評価作業を行う必要があるため、作業コストが高くなってしまう。そのため代替的な方法として、一つのシステムを基準となるシステムとして選択し、残りのシステムとそのシステムとの間で一対一の優劣評価を行うことで、その優劣の数によって順位付けし、結果として全システムの順位付けを行うことが可能となる。評価型ワークショップの WAT では、phrase-based SMT を基準システムとし、それと他の参加システムとの間の比較によって、全システムの順位付けを行っている[105][106]。２つの訳文を比較する際に、良いと判断するための指針は作業者に提示する必要はあるが、厳密な絶対評価に比べるとより短い時間での評価が可能であるため、評価コストを抑えることができると考えられる。

4.5 過去の精度調査

表題について、特許庁での調査、NTCIR での評価、WAT での評価、WMT での評価を記述する。

4.5.1 特許庁での調査

過去に特許庁で行われた下記の８つの調査について概要を記述する。

① 中国語機械翻訳技術に関する調査（H20）
② 特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査（H22）
③ 中国特許文献の和文抄録作成に対する機械翻訳の活用に関する調査（H23）
④ 中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査（H24）
⑤ 特許文献機械翻訳の品質評価手法に関する調査（H25）
⑥ 中国特許文献の機械翻訳のための辞書整備及び機械翻訳の品質評価に関する調査（H26）
⑦ 特許審査関連情報の日英機械翻訳文の品質評価に関する調査（H27）

⑧　中国特許文献の機械翻訳の品質評価及び辞書整備に関する調査（H27）

(1) 中国語機械翻訳技術に関する調査（H20）[122]

中国特許文献を対象とした中日機械翻訳の品質向上を目的とした調査研究である。翻訳対象文書は中国公開特許公報（中国語）であり、G セクション（物理）と H セクション（電気）を対象にしている。評価対象システムは規則方式の翻訳システムであり、特許翻訳用にチューニングする前（電気分野の専門用語辞書は利用）と後（ユーザ辞書利用（4,876 語)、ユーザ辞書＋定型パターン利用）の３システムである。試験文は、G セクション、H セクション各５件の公報から得られた 1,396 文であるが評価基準によっては一部の文に対してのみ評価している。

人手評価基準は、絶対評価として、正確さ（Adequacy)、流暢さ（Fluency）及び総合評価（１～５の５段階）の３種類、相対評価として、チューニング前と後での改善の割合を用いた。自動評価基準は、BLEU である。また、特許審査を想定した先行技術調査での有効性を測る目的達成評価も実施している。

人手評価（絶対評価）及び自動評価の評価結果の一部を、表４．５．１-１に示す（文献 [122] の表 3.2-5 と表 3.2-11 を改変)。ユーザ辞書や定型パターンを利用することで評価値が向上している。

表４．５．１-１　人手評価（絶対評価）及び自動評価の評価結果

	チューニング前	辞書利用	辞書+定型パターン利用
総合評価	2.49	2.85	2.94
流暢さ	2.74	2.91	3.05
正確さ	2.81	3.19	3.26
BLEU	6.74	6.76	8.26

人手評価（相対評価）の評価結果の一部を表４．５．１-２に示す（文献 [122] の表 3.2-7 と表 3.2-8 を改変)。表中、「前→辞書利用」とはチューニング前のシステムからユーザ辞書利用のシステムへの改善の割合（%）を意味し、「辞書→辞書+定型」とはユーザ辞書利用のシステムからユーザ辞書＋定型パターン利用のシステムへの改善の割合（%）を意味する。

表４．５．１-２　人手評価（相対評価）の評価結果

	前→辞書利用	辞書→辞書+定型
相対評価	53	16

[122] 特許庁:中国語機械翻訳技術に関する調査報告書、平成 21 年 3 月.

(2) 特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査（H22）[123]

表題の文書を対象にした日英機械翻訳の精度評価を行っている。評価対象システムは Advanced Industrial Property Network（AIPN）と市販の PAT-Transer（PATT）である。試験文は、日本国の公開特許公報より 206 件、日本国の拒絶理由通知書より 305 件を選択し、各文献から１文ずつを抽出した合計 511 文である。評価基準は、人手評価が、意味理解度、構文理解度、訳語評価の３基準、自動評価が、BLEU、NIST、IMPACT の３基準である。自動評価のための基準翻訳文として試験文の各文に対して２文ずつの英訳文を用意した。

評価結果の一部を表４．５．１-３に示す（文献 [123] の表 3.3.-1、表 3.4.-1、グラフ 4.2.2.1.-1、グラフ 4.2.2.1.-2、グラフ 4.2.2.1.-3、グラフ 4.2.2.2.-1、グラフ 4.2.2.2.-2、グラフ 4.2.2.2.-3、グラフ 4.2.2.3.-1、グラフ 4.2.2.3.-2、グラフ 4.2.2.3.-3 を改変)。表中、意味理解度の翻訳精度１と翻訳精度２については、文献 [123] の 19 ページを参照されたい。また、自動評価の single reference は基準翻訳文として１文を用いた場合、multi reference は２文を用いた場合である。いずれの基準でも AIPN のほうが PATT より評価値が高い。

同文献では自動評価結果と人手評価結果との相関も求めており、BLEU と NIST は人手評価結果との相関が低いことが指摘されている。IMPACT と人手評価結果の相関は 0.2 以上となる場合が多い。

表４．５．１-３　評価結果

評価規準	全データ		公開特許公報		拒絶理由通知書	
	AIPN	PATT	AIPN	PATT	AIPN	PATT
意味理解度(翻訳精度1)	62.08	50.00	63.35	49.64	61.23	50.25
意味理解度(翻訳精度2)	36.89	18.92	36.70	18.74	37.02	19.75
構文理解度(文あたり誤り数)	2.3	2.9	2.1	2.8	2.4	3.0
構文理解度(文節あたり誤り率 %)	10.4	13.5	9.5	12.8	11.0	13.9
訳語評価(文あたり誤り数)	1.0	1.2	1.2	1.4	0.9	1.0
訳語評価(文節あたり誤り率 %)	4.6	5.3	5.5	6.3	3.9	4.6
BLEU (single reference)	0.18	0.17	0.20	0.18	0.17	0.16
BLEU (multi reference:2)	0.26	0.23	0.28	0.24	0.25	0.23
NIST (single reference)	5.70	5.50	5.80	5.50	5.20	5.10
NIST (multi reference:2)	7.40	7.10	7.50	7.10	6.90	6.80
IMPACT (single reference)	0.32	0.30	0.33	0.30	0.31	0.29
IMPACT (multi reference:2)	0.35	0.33	0.36	0.33	0.33	0.32

[123] 特許庁:特許審査関連情報の機械翻訳による英語提供に対する精度評価に係る調査報告書,平成 23 年 2 月.

(3) 中国特許文献の和文抄録作成に対する機械翻訳の活用に関する調査（H23）[124]

本調査では、中国語のみで公開されている中国公開特許公報を対象に実際に和文抄録作成事業を開始する際の課題や問題点を洗い出し、かつ翻訳精度を維持しつつ効率的に事業を実施する方策としての機械翻訳の利用可能性等について基礎調査を行うことを目的に行われた。具体的には、以下の３種類の評価を行った。

① 中国公開特許公報から 200 件のサンプルを抽出し、それらをもとに機械翻訳利用を含む５種類の中国和文抄録サンプルデータを作成し、翻訳精度、検索精度、情報量の観点から５段階評価を行い、比較した。
② 40 件のサンプルに対して、４種類の異なる中日機械ソフト／サービスで公報全文の機械翻訳を行い、特に検索観点からの評価を行った。４種類の機械翻訳システムは、１つ（MT#2）が SMT（Statistical Machine Translation,統計的機械翻訳）であり、他の３つは RBMT（Rule-based Machine Translation,規則方式機械翻訳）である。
③ 第２の評価で、最も評価値の高かったシステム（MT#2）について中国和文抄録作成の評価を追加で行った。

まず、評価①における５種類の中国和文抄録サンプルデータ作成方法は以下のとおりである。

和抄#1：中国語公開特許公報の全文読解による人手和文抄録作成
和抄#2：ＣＰＡ（英文要約）の人手翻訳
和抄#3：ＣＰＡ（英文要約）の機械翻訳
和抄#4：中国語出願人要約の人手翻訳
和抄#5：中国語出願人要約の機械翻訳

評価①の結果を表４．５．１-４に、評価②の結果を表４．５．１-５に、評価③の結果を表４．５．１-６にそれぞれ示す（なお、表４．５．１-４には、評価③で行った追加評価（MT#2）も含めて示した）。人手翻訳は機械翻訳に比較して評価が高いこと、SMT である MT#2 の評価値が最も高いこと、MT#2 は CPA（英文要約）を経由した人手翻訳である和抄#2 より検索精度の点で評価値が高いことが示されている。

[124] 特許庁：中国特許文献の和文抄録作成に対する機械翻訳の活用に関する調査報告書（概要版）、平成 23 年 11 月.

表４．５．１-４　中国和文抄録サンプルデータの品質評価結果

評価の観点	和抄#1	和抄#2	和抄#3	和抄#4	和抄#5	MT#2
翻訳精度	4.9	4.6	3.1	4.8	2.4	2.4
検索精度	4.6	3.2	2.5	4.4	2.4	3.3
情報量	4.6	3.5	2.4	4.2	2.0	2.5
総合評価	14.1	11.3	8.0	13.4	6.8	8.2

表４．５．１-５　機械翻訳によるキーワード含有率の評価結果

	MT#1	MT#2	MT#3	MT#4
重要キーワード含有率 (%)	52	65	40	52

表４．５．１-６　機械翻訳データと中国和文抄録データの品質評価結果の対比

評価の観点	MT#2	和抄#1	和抄#2	和抄#3	和抄#4	和抄#5
翻訳精度	2.4	4.8	4.5	2.8	4.8	2.1
検索精度	3.3	4.4	2.9	2.1	4.0	1.9
情報量	2.5	4.4	3.2	2.3	4.0	1.9

(4) 中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査（H24）[125]

中日対訳コーパスを作成し、そこから中日対訳辞書データを作成するとともに、RBMT システムに対して、作成した辞書データを追加することによる翻訳精度変化を評価した。加えて、中日対訳コーパスの量の違いによる SMT の翻訳精度の変化を評価した。

まず、DOCDB に蓄積されているパテントファミリー情報（family-id）を利用して、特許庁から貸与された 2005 年～2009 年の中国公開特許公報（約 105 万件）と技術内容が対応する日本公開特許公報のリスト（約 27 万件）を作成し、それらパテントファミリーから対訳文を抽出した。抽出した対訳コーパス１は約 6,700 万文対である。その中から、中日辞書作成の目的で、対訳のスコアが高い文対を抽出した。得られた対訳コーパス２は約 1,552 万文対である。この対訳コーパス２から中日対訳辞書候補を抽出し、人手確認を経て 100 万語の中日対訳辞書データを作成した。

次に、RBMT システム（The 翻訳エンタープライズ V15）を対象に、中日対訳辞書データを追加する前のシステム（RBMT1）と追加した後のシステム（RBMT2）に対して、試験データ中に含まれる中国語名詞の翻訳結果の良否を評価基準として評価した。試験データは中国語特許文献の要約からサンプルした 160 件である。評価の結果を表４．５．１-７に示す。RBMT1 と比較して RBMT2 の正訳率は 17%向上している。

[125] 特許庁:中国特許文献の機械翻訳のための中日辞書整備及び機械翻訳性能向上に関する調査報告書（概要版）,平成 25 年 3 月.

表４．５．１-７　RBMT における辞書追加の効果

	RBMT1	RBMT2
正訳率 (%)	61	78

次に、SMT 訓練用の対訳コーパスを変化させて精度比較を行った。使用した対訳コーパスは前述した対訳コーパス１からスコア上位 2,500 万文対を抽出し、そこから 1,000 万文対をランダムに抽出した対訳コーパス３と同じ 2,500 万文対から 100 万文対をランダムに抽出した対訳コーパス４である。利用した SMT システムは NICT が開発したフレーズベースの SMT であり、対訳コーパス３（1,000 万文対）で訓練したシステムを SMT2、対訳コーパス４（100 万文対）で訓練したシステムを SMT1 とする。試験データ及び評価基準は前述した RBMT の評価と同一である。評価の結果を表４．５．１-８に示す。SMT1 と比較して SMT2 の正訳率は６%向上している。

表４．５．１-８　SMT における対訳コーパス数の効果

	SMT1	SMT2
正訳率 (%)	69	75

なお、RBMT と SMT ではシステムの条件が異なるため、表４．５．１-７と表４．５．１-８の結果を単純に比較することはできない。

(5) 特許文献機械翻訳の品質評価手法に関する調査（H25）[126]

特許庁における機械翻訳に関連する各種事業において機械翻訳結果の品質を適切に評価するための翻訳品質評価手法を調査・検証することを目的の一つとしており、「特許庁機械翻訳品質評価手法」を作成するとともに、英日機械翻訳を対象にして同手法を実践し、その妥当性を検証した。なお、文献[126]には機械翻訳品質評価に関する主要な先行研究（2.1.1 節、2.1.2.節）及び過去の特許庁調査における翻訳品質評価手法の内容（2.1.3 節）も記述されている。

「特許庁機械翻訳品質評価手法」は、内容伝達レベルの評価、重要技術用語の翻訳精度の評価、流暢さの評価、チェックリストに基づく評価、の４種類の基準を用いており、前二者を主要項目、後二者を補助項目としている。「内容伝達レベルの評価」は「正確さ（Adequacy)」の評価と類似しているが、正確さの度合いをパーセントで規定し、できるだけ客観的な評価ができるようにしている。「重要技術用語の翻訳精度の評価」においては「適訳語（A)」と「誤訳語（C)」の間に「可訳語（B)」を設定している。可訳語は「一般的に使われる訳語でなく同義の範囲外だが、正しい意味は分かる」という訳語であり、

126 日本特許情報機構:特許文献機械翻訳の品質評価手法に関する調査報告書,平成 26 年 2 月.

「検索」用途においては有用性が乏しいが、「粗読」用途や「精読」用途においては十分な有用性がある訳語である。

試験文は英日パテントファミリーから抽出された 199 文である。また、試験に用いた技術用語の総数は 524 語である。評価者は、英日特許翻訳者（すべての評価基準に対する評価)、機械翻訳専門家（内容伝達レベルの評価のみ)、和文抄録校閲者（内容伝達レベルの評価のみ）の３名である。評価対象システムは RBMT が３種類、SMT が２種類である。

評価結果の概要を表４．５．１-９に示す（文献 [126] の表 1-5、表 1-6、表 1-7、表 1-8 を改変)。いずれの評価基準でも SMT4 が最良と評価されている。ただし、本調査における品質評価実践は、「特許庁手法」の妥当性の検証を主目的とした試行評価であり、評価対象とした５種類の英日機械翻訳エンジンの精密な性能比較を目的としたものではない。したがって、本評価結果をそのまま各エンジンの性能比較に利用することは推奨されていない。

また、文献 [126] では、評価結果の差（チェックリストによる評価を除く）について符号検定を実施している（8.2.4 節、8.3.3 節、8.4.3 節)。その結果、200 文程度の試験文を用いた場合に「システム間に差があると認定するのに必要な評価値の差」が得られ、これも表４．５．１-９に示す。さらに文献 [126] では、内容伝達レベルの評価値と重要技術用語の翻訳精度の評価値を組み合わせて、各種実務用途への有用性を直接判断する「絶対評価スコア」を求める方法が提案されている（4.3.9 節、8.6 節)。また、パテントファミリーから抽出した試験文が評価結果に与える影響についても考察している（9.4 節)。

表４．５．１-９　評価結果の概要

	RBMT1	RBMT2	RBMT3	SMT4	SMT5	必要な評価値の差
内容伝達レベル	2.39	2.79	3.01	3.15	2.76	0.20
重要技術用語(適訳率　%)	52.1	60.5	66.4	84.5	75.2	6.0
重要技術用語(適訳＋可訳率　%)	72.7	77.1	83.6	92.3	84.0	6.0
流暢さ	2.46	2.78	3.02	3.16	2.83	0.20
チェックリスト(必須項目　カウント数)	735	583	505	411	560	

(6) 中国特許文献の機械翻訳のための辞書整備及び機械翻訳の品質評価に関する調査（H26）[127]

2013 年の中日パテントファミリーから対訳コーパスと 20 万語の中日対訳辞書データを作成するとともに、「特許文献機械翻訳の品質評価手順」の改訂を目的として、調査を実施した。「特許庁機械翻訳品質評価手法（特許文献機械翻訳の品質評価手順)」は英日翻訳を例として設定されたため、本調査では、中日翻訳における「内容伝達レベルの評価」の評価基準を、具体例を用いて示すとともに、チェックリストに基づく評価のための評価項目を設定した。

[127] 日本特許情報機構:中国特許文献の機械翻訳のための辞書整備及び機械翻訳の品質評価に関する調査, 平成 27 年 3 月.

評価対象システムは、１種類の RBMT に対して、ユーザ辞書を利用しないシステム（機械翻訳 B)、ユーザ辞書を利用したシステム（機械翻訳 S と機械翻訳 D）の３システムである。機械翻訳 S と機械翻訳 D の違いはユーザ辞書の訳語を文単位での出現頻度の最上位の訳語に定めたもの（S）と文献単位での出現頻度の最上位の訳語に定めたもの（D）の違いである。試験文数は 150 文である。機械翻訳 S と機械翻訳 D に違いがある 61 文に対して内容伝達レベルの評価を行った。機械翻訳 S と機械翻訳 D に違いがない 89 文に対しては、機械翻訳 S と機械翻訳 B に対して同様の評価を行った。

評価結果の一部を表４．５．１-１０に示す（文献 [127] の表 4.4-1、表 4.5-1 を改変)。機械翻訳 D は機械翻訳 S より内容伝達レベルの評価値が高いが、その差は小さい。一方、機械翻訳 S は機械翻訳 B より内容伝達レベルの評価値が高い。

表４．５．１-１０　内容伝達レベルの評価結果

(a) 機械翻訳 S と機械翻訳 D の比較（61 文）

内容伝達レベル	機械翻訳 S	機械翻訳 D
5	0	0
4	6	8
3	11	11
2	17	14
1	27	28
平均値	1.93	1.98

(b) 機械翻訳 S と機械翻訳 B の比較（89 文）

内容伝達レベル	機械翻訳 S	機械翻訳 B
5	0	0
4	13	7
3	21	20
2	26	18
1	29	44
平均値	2.20	1.89

(7) 特許審査関連情報の日英機械翻訳文の品質評価に関する調査（H27）[128]

本調査は、AIPN による日英機械翻訳について、平成 22 年度以降の調査で実施された精度向上策の評価を行うとともに、他の近年進歩が著しい機械翻訳サービスによる翻訳文との精度比較や、人手により作成された公開特許公報英文抄録（以下、「PAJ」という。）と比較した先行技術調査時の有用性を評価することで、AIPN を含む海外特許庁向けの情報提供サービスにおける日英翻訳の今後のあり方について検討するための基礎資料を得ることを目的としている。ここでは日英機械翻訳の翻訳精度の調査の部分について概要を記述する。

評価対象のシステムは、特許庁の機械翻訳（J-PlatPat)、Espacnet の機械翻訳（Google 翻訳)、NICT のみんなの自動翻訳@TexTra®（みんなの翻訳）の３種類である。試験文は、日本の公開特許公報 210 文献から抽出した 630 文と拒絶理由通知書 19 件から抽出した 70 文の合計 700 文である。公開特許公報は 35 の技術分野について各分野６文献ずつを選定した。各文献の要約・特許請求の範囲・発明な詳細な説明から各１文を選定した。また、公開特許公報から、１分野について 10 語の技術用語をその分野の重要技術用語として抽

[128] 日本特許情報機構:特許審査関連情報の日英機械翻訳文の品質評価に関する調査報告書,平成 28 年 3 月.

出した。人手評価基準は「内容伝達レベルの評価」及び「重要技術用語の翻訳精度の評価」である。評価者数は「内容伝達レベルの評価」では 2 名であり、「重要技術用語の翻訳精度の評価」では 10 名（合計）である。自動評価基準は、BLEU、NIST 及び RIBES である。

評価結果を表４．５．１－１１に示す（文献 [128] の図 2-5、図 2-15、図 2-18、図 2-19、図 2-20 から作成）。内容伝達レベルの評価値は、J-PlatPat>みんなの翻訳>Google 翻訳の順で評価値が高い。重要技術用語の翻訳精度は、適訳+可訳率ですべてのシステムで 95%以上となっている。拒絶理由通知書翻訳の自動評価値は、J-PlatPat>Google 翻訳>みんなの翻訳の順で評価値が高く、公開特許公報翻訳の自動評価値は、BLEU と NIST で、Google 翻訳>みんなの翻訳>J-PlatPat の順、RIBES では、みんなの翻訳>J-PlatPat>Google 翻訳の順で評価値が高かった。人手評価結果と自動評価結果には齟齬が見られる。

文献 [128] では、分野別の詳細な分析やパテントファミリーの有無による評価値の違いについての分析なども行われている。

表４．５．１－１１　評価結果

（a）内容伝達レベルの評価（公開特許公報）

内容伝達レベル	J-PlatPat	Google翻訳	みんなの翻訳
5 (%)	20	13	17
4以上 (%)	47	31	42
3以上 (%)	74	61	70
2以上 (%)	95	91	93
1以上 (%)	100	100	100
平均値	3.36	2.96	3.22

（b）内容伝達レベルの評価（拒絶理由通知書）

内容伝達レベル	J-PlatPat	Google翻訳	みんなの翻訳
5 (%)	7	1	0
4以上 (%)	29	5	6
3以上 (%)	52	29	33
2以上 (%)	83	67	68
1以上 (%)	100	100	100
平均値	2.71	2.02	2.07

（c）重要技術用語の翻訳精度の評価

重要技術用語	J-PlatPat	Google翻訳	みんなの翻訳
適訳率(%)	79	81	83
適訳+可訳率(%)	95	95	96

（d）自動評価（公開特許公報）

自動評価規準	J-PlatPat	Google翻訳	みんなの翻訳
BLEU	0.24	0.28	0.26
NIST	6.4	7.2	6.9
RIBES	0.71	0.68	0.74

（e）自動評価（拒絶理由通知書）

自動評価規準	J-PlatPat	Google翻訳	みんなの翻訳
BLEU	0.37	0.27	0.22
NIST	6.4	5.7	5.1
RIBES	0.73	0.69	0.65

(8) 中国特許文献の機械翻訳の品質評価及び辞書整備に関する調査（H27）[129]

本調査は、以下の３点を目的として実施した。

① 中国特許文献の中日機械翻訳文の人手による評価と自動評価をそれぞれ行うことで、中国特許文献の中日機械翻訳の技術分野ごとの精度を把握する。
② 機械翻訳の精度が低い分野において重点的に中日対訳辞書データを作成する。
③ 「特許文献機械翻訳の品質評価手順」についての改善すべき点を把握する。

ここでは、①について概要を記述する。評価対象文は中国特許文献の「発明の詳細な説明」に相当する部分から抽出した 700 文である（35 分野×20 文献×１文／文献）。この際、日本語のパテントファミリーのある文献は技術分野ごとに半数未満にした。基準翻訳文は日中双方ネイティブレベルかつ、特許文献の翻訳経験がある作業者が、各自得意な技術分野を担当して作成した。

評価対象システムは「中韓文献翻訳・検索システム」及び「みんなの翻訳」サービスである。人手評価基準は「内容伝達レベル」及び「重要技術用語の翻訳精度」であり、自動評価基準は BLEU と RIBES である。評価者数は「内容伝達レベル」が２チーム、「重要技術用語の翻訳精度」が２名である。

700 文全体の評価結果を、表４．５．１-１２に示す（文献 [129] の図 3.2.1-1、図 3.2.2-1、図 3.4.1-1、図 3.4.2-1 から作成）。特許庁翻訳（中韓文献翻訳・検索システム）よりみんなの翻訳の方が、評価値が高いことが示されている。

[129] 高電社:中国特許文献の機械翻訳の品質評価及び辞書整備に関する調査報告書,平成 28 年 2 月.

表４．５．１-１２　評価結果（700 文全体）

（a）評価値の平均値（重要技術用語の翻訳精度は%）

評価規準	特許庁翻訳	みんなの翻訳
内容伝達レベル	3.15	3.87
重要技術用語の翻訳精度 (適訳率 %)	69.80	82.70
重要技術用語の翻訳精度 (適訳+可訳率 %)	79.90	88.70
BLEU	20.20	25.93
RIBES	0.7724	0.8022

（b）標準偏差

評価規準	特許庁翻訳	みんなの翻訳
内容伝達レベル	1.03	0.97
BLEU	13.97	16.75
RIBES	0.0942	0.0939

4.5.2　NTCIR での評価

NTCIR（NII Testbeds and Community for Information access Research）は情報アクセス技術に関する国際プロジェクトであり、特許文書を対象にした機械翻訳についてのワークショップが 2008 年から 2013 年にかけて NTCIR-7～NTCIR-10 の４回行われた。

(1) NTCIR-7（2008 年）[130]

2008 年に実施された、英語と日本語間の機械翻訳評価ワークショップであり、日米のパテントファミリーから抽出した約 180 万文対の特許文を訓練データとして配布された。

自動評価用の試験データは基準翻訳文が１文のもの single-reference（SRB）が 1,381 文対である。multi-reference は基準翻訳文が複数用意されたもので、英語側の基準翻訳文が３文のもの（S600）が 600 文対、２文のもの（S300）が 300 文対である。この３文または２文の基準翻訳文は異なる翻訳者による手翻訳で独立に翻訳されたが、S600 の手翻訳の一部は規則方式機械翻訳を下訳として用いている。なお、本来は基準翻訳文の作成に機械翻訳を下訳として用いることは適当でない。試験データの日本語側は、S300 は S600 に含まれ、S600 は SRB に含まれている。人手評価用の試験データは 100 文である。

評価基準は自動評価が BLEU、人手評価が Adequacy と Fluency である（他に検索観点からの評価も行われたが省略する）。S600 と S300 については BLEU を multi-reference BLEU として求めている。人手評価は３人の専門家によって行われた。

130 Atsushi Fujii, Masao Utiyama, Mikio Yamamoto and Takehito Utsuro : Overview of the Patent Translation Task at the NTCIR-7 Workshop, Proceedings of NTCIR-7 Workshop Meeting, Pages 389-400, December 16–19, 2008, Tokyo, Japan.

評価結果を翻訳方向別に表４．５．２-１に示す（文献 [130] の Table 2、Table 3 を改変)。各評価基準での上位３システムの結果である[131]。表中、Method 列は翻訳方式を示す。日英の結果について以下のことが分かる。single-reference（SRB）では自動評価値（BLEU）の上位はすべて SMT であるが、人手評価値（Adequacy と Fluency）の上位は RBMT であり評価が食い違っている。しかし、multi-reference（S300 と S600）にすることで RBMT の BLEU 値が上がった。例えば、single-reference では BLEU 値が低かった tsbmt が上位に上がっており、Adequacy と BLEU（SRB）の評価値の齟齬が小さくなっている。一方、人手評価の Fluency で１位の Japio は multi-reference の BLEU（S300）と BLEU（S600）においても上位に上がることはなく、齟齬は解消されていない。また、総じて、multi-reference にすることで BLEU 値が人手評価に近づくことが見える。

人手評価の結果は、表４．５．２-１の評価結果に示すように、日英翻訳においては明確に RBMT の評価が高く、英日翻訳においては未だ RBMT の評価が高いが SMT が肉薄してきている。

[131] 各チームのシステム中で最高評価のシステムに対する評価値を示した。本文の以下の記述でも同様の方法で示した。

表４．５．２-１　評価結果

（a）日→英

System	Method	BLEU (SRB)
NTT	SMT	27.20
Moses	SMT	27.14
MIT	SMT	27.14

System	Method	BLEU (S300)
tsbmt	RBMT	37.51
MIT	SMT	37.31
Moses	SMT	36.02

System	Method	BLEU (S600)
tsbmt	RBMT	48.02
MIT	SMT	44.69
NTT	SMT	43.72

System	Method	Adequacy
tsbmt	RBMT	3.81
JAPIO	RBMT	3.71
MIT	SMT	3.15

System	Method	Fluency
JAPIO	RBMT	4.02
tsbmt	RBMT	3.94
MIT	SMT	3.66

（b）英→日

System	Method	BLEU (SRB)
Moses	SMT	30.58
HCRL	SMT	29.97
NiCT-ATR	SMT	29.15

System	Method	Adequacy
tsbmt	RBMT	3.53
Moses	SMT	2.90
NTT	SMT	2.74

System	Method	Fluency
Moses	SMT	3.69
tsbmt	RBMT	3.67
NTT	SMT	3.54

(2) NTCIR-8（2009 年）[132]

2009 年に実施された、英語と日本語間の機械翻訳評価ワークショップであり、日米のパテントファミリーから抽出した約 320 万文対の特許文を訓練データとして用いた。

試験データは日英方向が 1,251 文対であり、英日方向が 1,119 文対である。いずれも、基準翻訳文数は１である（single reference)。評価基準は BLEU のみであり、人手評価は実施していない。評価結果を表４．５．２-２に示す（文献 [132] の Table 1、Table 2 を改変)。NTCIR-7 の結果と比較して、日英、英日ともに BLEU 値が上昇している。翻訳方式（Method）の中で Hybrid とは複数の翻訳方式を組み合わせたものである。

[132] Atsushi Fujii, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya and Sayori Shimohata : Overview of the Patent Translation Task at the NTCIR-8 Workshop, Proceedings of NTCIR-8 Workshop Meeting, Pages 371-376, June 15-18, 2010, Tokyo, Japan.

表４．５．２-２　評価結果

（a）日→英

System	Method	BLEU
EIWA	Hybrid	34.30
NICT	SMT	30.32
Moses	SMT	29.08

（b）英→日

System	Method	BLEU
NICT	SMT	35.87
Moses	SMT	35.27
DCU	Hybrid	33.03

(3) NTCIR-9（2011 年）[133]

2011 年に実施された、日英・英日方向及び中英方向の機械翻訳評価ワークショップである。日英・英日方向の訓練データは NTCIR-8 と同一の日米のパテントファミリーから抽出した約 320 万文対の特許文を用いた。中英方向の訓練データは中国（香港）の組織が作成した約 100 万文対の特許文を用いた。さらに単言語コーパスとして、英語と日本語の 13 年分のデータが提供された。

自動評価の試験データはすべての方向について 2,000 文対のデータを用いた。自動評価の評価基準は BLEU、NIST、RIBES である。人手評価の試験データはすべての方向について 300 文対である。人手評価の評価基準は Adequacy と Acceptability である。評価者数は３名であるが、一人の評価者は 100 文を評価しており、異なる評価者は異なる文を評価しているため、合計の評価文数は 300 文となる。評価者には基準翻訳文を開示していない。評価結果を表４．５．２-３に示す（文献 [133] の Table 5、Table 7、Table 9、Table 11、Table 13、Table 15、Table 21、Table 22、Table 23 を改変）。表中、Acceptability の評価値は Pairwise score である。

日英方向の人手評価値は、RBMT あるいは Hybrid が上位である一方、自動評価では Hybrid 又は SMT が上位であり、齟齬が見られる。英日方向では NTT-UT（SMT）がすべての評価基準で１位となっているが、人手評価値では RBMT が上位にある。

中英方向では、BBN（SMT）がすべての評価基準に対して１位にある。

[133] Isao Goto, Bin Lu, Ka Po Chow, Eiichiro Sumita and Benjamin K. Tsou : Overview of the Patent Machine Translation Task at the NTCIR-9 Workshop, Proceedings of NTCIR-9 Workshop Meeting, Pages 559-578, December 6-9, 2011, Tokyo, Japan.

表４．５．２-３　評価結果

（a）日→英

System	Method	Adequacy
JAPIO	RBMT	3.667
RBMT1	RBMT	3.530
EIWA	Hybrid	3.430

System	Method	Acceptablility
JAPIO	RBMT	0.712
RBMT1	RBMT	0.674
EIWA	Hybrid	0.638

System	Method	BLEU
EIWA	Hybrid	31.69
RWTH	SMT	30.32
KLE	SMT	29.55

System	Method	NIST
RWTH	SMT	7.879
KLE	SMT	7.828
EIWA	Hybrid	7.816

System	Method	RIBES
EIWA	Hybrid	0.7404
NAIST	SMT	0.7307
NTT-UT	SMT	0.7195

（b）英→日

System	Method	Adequacy
NTT-UT	SMT	3.670
RBMT6	RBMT	3.507
JAPIO	RBMT	3.463

System	Method	Acceptablility
NTT-UT	SMT	0.695
RBMT6	RBMT	0.656
JAPIO	RBMT	0.652

System	Method	BLEU
NTT-UT	SMT	39.48
KLE	SMT	35.10
ICT	SMT	32.91

System	Method	NIST
NTT-UT	SMT	8.713
KLE	SMT	8.285
ICT	SMT	8.170

System	Method	RIBES
NTT-UT	SMT	0.7813
TORI	Hybrid	0.7479
KLE	SMT	0.7429

（c）中→英

人手評価

System	Method	Adequacy
BBN	SMT	4.033
NEU	SMT	3.510
RWTH	SMT	3.420

System	Method	Acceptability
BBN	SMT	0.744
RWTH	SMT	0.546
NEU	SMT	0.544

自動評価

System	Method	BLEU
BBN	SMT	39.44
IBM	SMT	36.11
RWTH	SMT	35.69

System	Method	NIST
BBN	SMT	8.911
RWTH	SMT	8.629
IBM	SMT	8.509

System	Method	RIBES
BBN	SMT	0.8327
IBM	SMT	0.7972
RWTH	SMT	0.7961

(4) NTCIR-10（2013 年）[134]

2013 年に実施された、日英・英日方向及び中英方向の機械翻訳評価ワークショップである。日英・英日方向の訓練データは NTCIR-8 と同一の日米のパテントファミリーから抽出した約 320 万文対の特許文を用いた。中英方向の訓練データは中国（香港）の組織が作成した約 100 万文対の特許文を用いた。さらに単言語コーパスとして、英語と日本語の 13 年分のデータが提供された。

自動評価の試験データは日英・英日方向及び中英方向すべてについて 2,300 文対のデータを用いた。自動評価の評価基準は BLEU、NIST、RIBES である。人手評価の試験データはすべての方向について 300 文対である。人手評価の評価基準は Adequacy と Acceptability である。評価者数は３名である。一人の評価者は 100 文を評価しており、異なる評価者は異なる文を評価しているため、合計の評価文数は 300 文となる。評価者には基準翻訳文を開示していない。

NTCIR-10 のユニークな人手評価基準として、実際の特許審査における有効性を見るための評価を行った（PEE:Patent Examination Evaluation)。評価値は６段階である（文献 [134] の Table 2 を参照)。これは中英と日英の方向のみ実施し、評価者は英語に堪能な審査経験者であり、評価者数は２名である。

評価結果を表４．５．２-４に示す（文献 [134] の Table 7、Table 9、Table 11、Table 13、Table 15、Table 17、Table 19、Table 20、Table 28、Table 29、Table 30 を改変)。Acceptability の評価値は Pairwise score である。PEE の評価値は６段階中上位２段階の占める割合（%）を示した。

日英の人手評価は RBMT が上位である。特に Japio（RBMT）の PEE は 100%を達成している。一方、日英の自動評価では SMT 又は Hybrid が上位であり、人手評価との齟齬が見られる。一方、英日では NTITI（SMT）がすべての評価基準で１位であり、中英では BBN（SMT）がすべての評価基準で１位である。

[134] Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita and Benjamin K. Tsou : Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop, Proceedings of the 10th NTCIR Conference, Pages 260-286, June 18-21, 2013, Tokyo, Japan.

表４．５．２-４　評価結果

（a）日→英

System	Method	Adequacy
JAPIO	RBMT	3.67
RBMT1	RBMT	3.57
EIWA	Hybrid	3.53

System	Method	Acceptablility
JAPIO	RBMT	0.630
TORI	Hybrid	0.580
EIWA	Hybrid	0.567

System	Method	PEE
JAPIO	RBMT	100
EIWA	Hybrid	79
NTITI	SMT	64

System	Method	BLEU
RWTH	SMT	33.77
NTITI	SMT	32.55
EIWA	Hybrid	32.50

System	Method	NIST
RWTH	SMT	8.550
HDU	SMT	8.442
EIWA	Hybrid	8.270

System	Method	RIBES
EIWA	Hybrid	0.7402
NTITI	SMT	0.7324
JAPIO	RBMT	0.7214

（b）英→日

System	Method	Adequacy
NTITI	SMT	3.84
JAPIO	RBMT	3.53
RBMT6	RBMT	3.47

System	Method	Acceptablility
NTITI	SMT	0.659
EIWA	Hybrid	0.568
JAPIO	RBMT	0.560

System	Method	BLEU
NTITI	SMT	42.89
EIWA	Hybrid	36.93
BJTUX	SMT	34.45

System	Method	NIST
NTITI	SMT	9.265
EIWA	Hybrid	8.501
BJTUX	SMT	8.421

System	Method	RIBES
NTITI	SMT	0.7984
EIWA	Hybrid	0.7692
TSUKU	SMT	0.7566

（c）中→英

人手評価

System	Method	Adequacy
BBN	SMT	4.15
RWSYS	Hybrid	3.52
SRI	SMT	3.51

System	Method	Acceptability
BBN	SMT	0.685
RWTH	SMT	0.533
RWSYS	Hybrid	0.523

System	Method	PEE
BBN	SMT	88
RWSYS	Hybrid	36
SRI	SMT	7

自動評価

System	Method	BLEU
BBN	SMT	42.68
RWSYS	Hybrid	40.06
RWTH	SMT	39.75

System	Method	NIST
BBN	SMT	9.561
RWSYS	Hybrid	9.344
RWTH	SMT	9.299

System	Method	RIBES
BBN	SMT	0.8331
RWTH	SMT	0.8000
RWSYS	Hybrid	0.7987

4.5.3 WAT 及び WMT の評価

WAT（Workshop on Asian Translation）はアジア言語の機械翻訳に関するワークショップであり、2014 年から毎年行われている。対象文書は科学技術文献と特許文献である[135]。

WMT（Workshop on statistical Machine Translation）は欧州で行われている機械翻訳の評価ワークショップであり 2006 年から毎年開かれている。

(1) WAT2014[136]

2014 年に実施された英日・日英方向及び中日・日中方向の機械翻訳評価ワークショップである。対象文書は科学技術文献であり ASPEC（Asian Scientific Paper Excerpt Corpus）から抽出されている。訓練データは英日・日英が 300 万文対、中日・日中が 67 万文対である。試験データは英日・日英が 1,812 文対、中日・日中が 2,107 文対である。

自動評価基準は BLEU と RIBES であり、人手評価基準は対象システムとベースラインシステム[137]との対比較であり、クラウドソーシングによる評価である。人手評価の文数は 400 文であり、評価者数は３名であり、多数決により各評価文の評価値を定めた。最終的なシステムの人手評価値（HUMAN）は、当該システムがベースラインシステムと比較して良かった文数を W、悪かった文数を L、同等であった文数を T として、

$$\mathrm{HUMAN} = 100 \times \frac{W - L}{W + L + T}$$

で求めた。

評価結果を表４．５．３-１に示す（文献[136]の Table 11、Table 12、Table 13、Table 14 を改変）。表中、Method 列の EBMT とは用例方式機械翻訳（Example-based Machine Translation）を意味する。日英・英日、日中・中日すべてで NAIST（SMT）が１位である。多くの場合で、SMT が他の方式より良い評価値を得ている。

---

135 2016 年度からは文献種別としてニュース記事なども加わる予定である。言語種別にもヒンディ語やインドネシア語が加わる予定である。

136 Toshiaki Nakazawa, Hideya Mino, Isao Goto, Sadao Kurohashi and Eiichiro Sumita : Overview of the 1st Workshop on Asian Translation, Proceedings of the 1st Workshop on Asian Translation（WAT2014）, Pages 1-19, Tokyo, Japan, 4th October 2014.

137 Moses の Phrase-based SMT をベースラインとした。

表４．５．３-１　評価結果（WAT2014）

（a）日→英

System	Method	HUMAN
NAIST	SMT	40.50
SMT S2T	SMT	25.50
Kyoto-U	EBMT	25.00

System	Method	BLEU
NAIST	SMT	23.82
Kyoto-U	EBMT	21.07
TOSHIBA	Hybrid	20.61

System	Method	RIBES
NAIST	SMT	0.7235
TOSHIBA	Hybrid	0.7079
EIWA	Hybrid	0.7067

（b）英→日

System	Method	HUMAN
NAIST	SMT	56.25
WEBLIO-EJ1	SMT	43.25
Online A	Other	42.50

System	Method	BLEU
NAIST	SMT	35.03
WEBLIO-EJ1	SMT	32.69
Kyoto-U	EBMT	31.09

System	Method	RIBES
NAIST	SMT	0.8017
WEBLIO-EJ1	SMT	0.7850
Kyoto-U	EBMT	0.7664

（c）日→中

System	Method	HUMAN
NAIST	SMT	17.75
SMT S2T	SMT	14.00
Sense	SMT	10.00

System	Method	BLEU
NAIST	SMT	30.53
SMT S2T	SMT	28.65
NICT	SMT	27.98

System	Method	RIBES
NAIST	SMT	0.8296
SMT Hiero	SMT	0.8091
SMT S2T	SMT	0.8076

（d）中→日

System	Method	HUMAN
NAIST	SMT	50.75
SAS_MT	SMT	22.50
SMT T2S	SMT	16.00

System	Method	BLEU
NAIST	SMT	40.21
SAS_MT	SMT	37.42
SMT T2S	SMT	36.52

System	Method	RIBES
NAIST	SMT	0.8455
SAS_MT	SMT	0.8342
SMT T2S	SMT	0.8253

(2) WAT2015[138]

2015 年に実施された機械翻訳評価ワークショップであり、科学技術文献を対象にした英日・日英方向及び中日・日中方向の翻訳に加えて、特許文書を対象にした中日及び韓日方向の翻訳が加わった。科学技術文献は WAT2014 と同様の ASPEC（Asian Scientific Paper Excerpt Corpus）コーパスを用いて、訓練データ、試験データ共に WAT2014 と同一である。特許文書はパテントファミリーから抽出されたデータ（JPC : JPO Patent Corpus）を用い、訓練データは中日、韓日ともに 100 万文対、試験データは中日、韓日ともに 2,000 文である。自動評価基準は BLEU と RIBES であり、人手評価基準は対象システムとベ

[138] Toshiaki Nakazawa, Hideya Mino, Isao Goto, Graham Neubig, Sadao Kurohashi and Eiichiro Sumita, Overview of the 2nd Workshop on Asian Translation, Proceedings of the 2nd Workshop on Asian Translation（WAT2015）, Pages 1-28, Kyoto, Japan, 16th October 2015.

ースラインシステムとの対比較であり、クラウドソーシングによる評価である。人手評価の文数は 400 文であり、評価者数は５名であり、多数決の結果、差が２ポイント以上ある場合に良否を判断し、２ポイント未満の場合は同等として、各評価文の評価値を定めた。最終的なシステムの人手評価値（HUMAN）は WAT2014 と同一である。

クラウドソーシングの人手評価において上位３位以内に入ったシステムに対しては、特許庁が定めた「内容伝達レベルの評価」（JPO Adequacy）も行った。内容伝達レベルの評価文数は 200 文であり、評価者数は 2 名である。

評価結果を表４．５．３-２、表４．５．３-３、表４．５．３-４に示す（文献[138]の Table 14、Table 15、Table 16、Table 17、Table 18、Table 19、Figure 9 を改変）。ASPEC に対しては、日中方向の HUMAN を除いてすべての翻訳方向及び評価基準に対して、NAIST（SMT）が１位である。日中方向の HUMAN では TOSHIBA（Hybrid）が１位である。JPC に対しては、中日方向で EBMT や Hybrid が上位であり、韓日方向では SMT が上位である。

表４．５．３-２　評価結果（WAT2015 ASPEC）

（a）日→英

System	Method	HUMAN
NAIST	SMT	35.50
Kyoto-U	EBMT	32.50
TOSHIBA	Hybrid	25.00

System	Method	JPO adequacy
NAIST	SMT	3.83
Kyoto-U	EBMT	3.65
TOSHIBA	Hybrid	3.60

System	Method	BLEU
NAIST	SMT	25.41
TOSHIBA	Hybrid	23.00
Kyoto-U	EBMT	22.89

System	Method	RIBES
NAIST	SMT	0.7496
Kyoto-U	EBMT	0.7246
TOSHIBA	Hybrid	0.7185

（b）英→日

System	Method	HUMAN
NAIST	SMT	62.25
WEBLIO MT	SMT	53.75
naver	SMT	53.25

System	Method	JPO adequacy
NAIST	SMT	4.04
naver	SMT	4.00
WEBLIO MT	SMT	3.81

System	Method	BLEU
NAIST	SMT	35.83
naver	SMT	34.60
WEBLIO MT	SMT	33.23

System	Method	RIBES
NAIST	SMT	0.8114
naver	SMT	0.8073
WEBLIO MT	SMT	0.8047

表４．５．３-３　評価結果（WAT2015 ASPEC）

（c）日→中

System	Method	HUMAN
TOSHIBA	Hybrid	17.00
Kyoto-U	EBMT	16.00
SMT StoT	Other	7.75

System	Method	JPO adequacy
NAIST	SMT	3.17
Kyoto-U	EBMT	2.87
TOSHIBA	Hybrid	2.75

System	Method	BLEU
NAIST	SMT	31.61
Kyoto-U	EBMT	31.40
TOSHIBA	Hybrid	30.17

System	Method	RIBES
NAIST	SMT	0.8328
Kyoto-U	EBMT	0.8270
TOSHIBA	Hybrid	0.8173

（d）中→日

System	Method	HUMAN
NAIST	SMT	35.75
EHR	Hybrid	25.75
Kyoto-U	EBMT	18.50

System	Method	JPO adequacy
NAIST	SMT	3.88
Kyoto-U	EBMT	3.74
EHR	Hybrid	3.25

System	Method	BLEU
NAIST	SMT	41.75
EHR	Hybrid	39.43
Kyoto-U	EBMT	38.53

System	Method	RIBES
NAIST	SMT	0.8551
Kyoto-U	EBMT	0.8407
EHR	Hybrid	0.8377

表４．５．３-４　評価結果（WAT2015 JPC）

（a）中→日

System	Method	HUMAN
Kyoto-U	EBMT	27.50
TOSHIBA	Hybrid	24.25
EHR	Hybrid	22.00

System	Method	JPO adequacy
Kyoto-U	EBMT	3.41
EHR	Hybrid	3.32
TOSHIBA	Hybrid	3.25

System	Method	BLEU
TOSHIBA	Hybrid	41.82
Kyoto-U	EBMT	41.35
EHR	Hybrid	41.06

System	Method	RIBES
Kyoto-U	EBMT	0.8285
EHR	Hybrid	0.8270
ntt	SMT	0.8234

（b）韓→日

System	Method	HUMAN
Online A	Other	38.75
naver	SMT	14.75
NICT	SMT	10.50

System	Method	JPO adequacy
naver	SMT	4.78
NICT	SMT	4.76
Online A	Other	4.53

System	Method	BLEU
Sense	SMT	85.24
naver	SMT	71.38
TOSHIBA	SMT	71.01

System	Method	RIBES
Sense	SMT	0.9545
naver	SMT	0.9439
TOSHIBA	SMT	0.9437

(3) WMT での評価[139] [140] [141] [142] [143] [144] [145] [146] [147] [148]

WMT（Workshop on statistical Machine Translation）は欧州で行われている機械翻訳の評価ワークショップであり、2006 年から毎年開かれている。翻訳タスクが主タスクであるが、自動評価タスクなどいくつかのタスクが行われている。翻訳対象言語は主として欧州の言語（フランス語、ドイツ語、スペイン語、ロシア語など）であるが、2014 年では、ヒンディ語が加わっている。翻訳方向は英語から諸言語及びその逆である。翻訳対象文書は、EU 議会の文書（EUROPARL)、ニュース文書、Web から収集した文書などである。翻訳タスクの公式評価基準は人手評価である。2007 年までは Adequacy と Fluency による絶対評価を用いていたが、評価の信頼性が十分でないとして、その後は、システム間の総合的な順位付けによる相対評価に切り替えている（文献 [140] の６節と文献 [141] の３節参照)。WMT では、WAT と違いベースラインシステムを決めていないため、WAT の HUMAN に相当するスコアは得られず、参加システム間の順位とスコア（勝率）しか得られない。

---

139 Philipp Koehn and Christof Monz : Manual and Automatic Evaluation of Machine Translation between European Languages, Proceedings of the Workshop on Statistical Machine Translation, pages 102-121, New York City USA, June 2006.

140 Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder :（Meta-）Evaluation of Machine Translation, Proceedings of the Second Workshop on Statistical Machine Translation, pages 136–158, Prague Czech Republic, June 2007.

141 Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz and Josh Schroeder : Further Meta-Evaluation of Machine Translation, Proceedings of the Third Workshop on Statistical Machine Translation, pages 70–106, Columbus Ohio, June 2008.

142 Chris Callison-Burch, Philipp Koehn, Christof Monz and Josh Schroeder : Findings, Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 1–28, Athens Greece, March 2009.

143 Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki and Omar Zaidan : Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation, Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, pages 17-53, Uppsala Sweden, July 2010.

144 Chris Callison-Burch, Philipp Koehn, Christof Monz and Omar Zaidan : Findings of the 2011 Workshop on Statistical Machine Translation, Proceedings of the Sixth Workshop on Statistical Machine Translation, pages 22-64, Edinburgh Scotland, July 2011.

145 Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia : Findings of the 2012 Workshop on Statistical Machine Translation, Proceedings of the Seventh Workshop on Statistical Machine Translation, pages 10-51, Montreal Canada, June 2012.

146 Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia : Findings of the 2013 Workshop on Statistical Machine Translation, Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 1-44, Sofia Bulgaria, August 2013.

147 Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia and Aleš Tamchyna : Findings of the 2014 Workshop on Statistical Machine Translation, Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 12-58, Baltimore Maryland USA, June 2014.

148 Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia and Marco Turchi : Findings of the 2015 Workshop on Statistical Machine Translation, Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 1-46, Lisbon Portugal, September 2015.

5. アプリケーションとの連携

本節では、特許関連の文書やテキストを翻訳するサービスやアプリケーションを提供したり、他のアプリケーションと機能連携したりする際に考慮すべき課題を、代表的なアプリケーションや連携機能を例に説明する。

5.1 構造化文書の翻訳

5.1.1 構造化文書の翻訳におけるタグ復元処理

(1) タグ復元処理の目的

現在、特許庁から公開される特許公開公報をはじめとする特許情報の多くは、そのデータ格納形式として XML が用いられている。また、独立行政法人工業所有権情報・研修館がインターネットを通じて一般ユーザに提供しているデータは、パソコンのブラウザソフトで閲覧できるよう、最終的には HTML データとして提供されている。これらのデータにおいては、テキストに文書の論理構造やレイアウトに関する情報がタグの形で付与されているため、これらのデータの翻訳においては、その結果においても、原文文書で指定されている論理構造やレイアウトが可能な限り保たれていることが望ましい。

例えば、図５．１．１-１は HTML 文書の一例であり、図５．１．１-２はその文書をブラウザで表示した例である。図５．１．１-２の表示例から分かるとおり、テキストには font タグを使って色（図では赤）が付与されている単語や、クリックすると他のページの情報が参照できるようにしたリンク情報を設定する a タグが付与されている。また、文書のタイトルには h タグが付与されており、本文よりも大きなサイズで表示されている。例に示したこの英語文書を日本語に翻訳した結果としては図５．１．１-３に示すようになるのが望ましく、そのためには、図５．１．１-４に示すように、英語のテキストに付与されていたタグが、日本語で対応する表現に付与されなければならない。このように原文文書に付与されていたタグを原文で付与されていたのと同じように訳文に付与することを、ここではタグの復元と呼ぶ。

```
<html>
<body>
<h3>1.Introduction</h3>
We develop a <font color="red">machine translation</font> system, whose
architecture is outlined in <a href="./fig1.gif">Fig. 1</a>.
</body>
</html>
```

図５．１．１-１　HTML 文書の一例

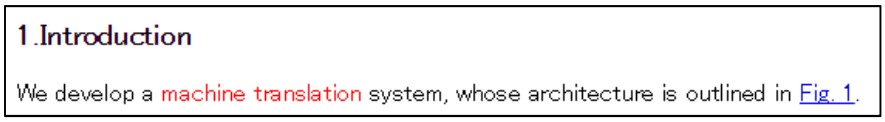


図５．１．１-２　HTML 文書のブラウザでの表示例

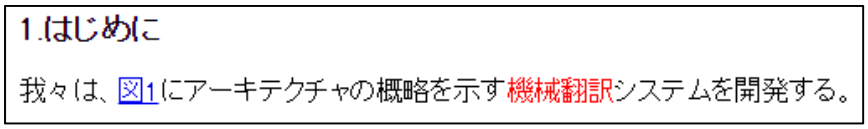


図５．１．１-３　HTML 文書の翻訳結果の表示例

```
<html>
<body>
<h3>1.はじめに</h3>
我々は、<a href="./fig1.gif">図 1</a>にアーキテクチャの概略を示す<font
color="red">機械翻訳</font>システムを開発する。
</body>
</html>
```

図５．１．１-４　HTML 文書の翻訳結果

(2) タグ復元を伴う翻訳処理方式の概要と課題

XML 文書の翻訳におけるタグ復元処理も、基本的には HTML 文書の場合と同じであるので、以下では HMTL 文書の翻訳を例にタグ復元を伴う処理について述べる。当該処理は基本的に次のステップで行われる。

・タグと、翻訳対象となるタグ以外のテキストを分離する。
・テキストを翻訳する。
・翻訳結果を原文のあった箇所に戻す。

しかし、5.1.1 の例で示したように、タグの中には、テキストに対するメタな情報を付与するために使われているタグもあるため、常にタグのある箇所ごとにテキストを区切って翻訳すると、翻訳が不適切なものとなってしまう。したがって、一文として翻訳すべきテキストを抽出する際には、そのようなタグに含まれる範囲を含める場合がある。このような文字を修飾するようなタグをここでは文内タグと呼ぶ。この文内タグが付与されている原文中の単語や表現は、翻訳によって目標言語の単語や表現へと変化するため、元の単語や表現とは異なる部分にタグを付与しなおす必要があるが、これが翻訳におけるタグ復元の本質的な問題となる。

文内タグの復元を行うには、基本的に以下の２つの方法が考えられる。

① 文内タグに含まれる範囲を、原文での表現のまま翻訳されずに訳文に出力されるようなシンボル（以下アンカーと呼ぶ）に変えて原文テキストに埋め込み、訳文にそのまま現れたアンカーを元のタグ情報に戻す方法。
② 原文から一旦タグを削除して翻訳を行い、翻訳結果にタグを戻す方法。タグが付与されていた単語や表現の位置情報を記録しておき、翻訳エンジンが出力する原文と訳文での単語の対応情報を利用して、原文でタグがついていたと思われる訳語にタグを付与する。

①の方法は、原文に出現したタグがほとんどの場合訳文に確実に復元できるというメリットがある。埋め込むアンカー文字列を、数字や未知語として処理される人工的な文字列にすることで、どのような翻訳方式であってもほとんどの場合は訳文に出力されるからである。その一方、原文に本来現れない文字列が翻訳処理されるテキストに埋め込まれることになるため、原文が非文法的になる恐れがある。例えば、RBMT では、正しく構文解析できなくなったり、訳語選択の規則が効かなくなったりする可能性がある。また、SMT では、翻訳モデルで定義された本来利用されるはずのフレーズが適用されなかったり、言語モデルでの評価結果に影響がでたりする場合があり、RBMT と同様に、タグがない場合とは異なる翻訳結果が得られる可能性がある。このように、いずれの場合にも、本来原文にはなかったシンボルを埋め込むことになるため、翻訳精度が低下する恐れがある。

②の方法では、原文はオリジナルのテキストのままであるため、①の方法で問題となる翻訳精度が低下する問題は生じない。しかし、逆に翻訳に起因する問題により適切にタグが復元できない場合もある。例えば、図５．１．１-５の英語と日本語の例では、英語原文の冠詞 a は日本語には訳出されない。そのため、もし英語の a にタグがついていても、それは本質的に復元しえないタグとなる。また、図５．１．１-５の例では、英語の machine translation という２単語が１つの専門用語として解析されたことが暗に示されているが、このときもし machine と translation の２つの語それぞれに異なるタグが付与されていると、訳文でも一まとまりの訳語となっている「機械翻訳」という語を「機械」と「翻訳」に分解して２つのタグを復元することは難しいため、結果として２つのタグのうちどちらかが復元されない可能性がある。

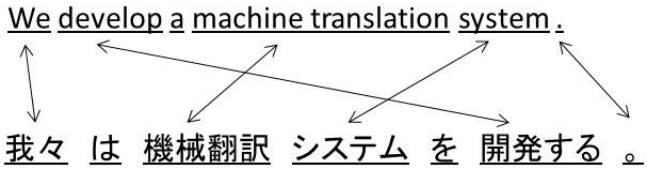


図５．１．１-５　原文と訳文の単語対応例

なお、タグの復元は、翻訳エンジンの外側で実現できる機能である。上で述べたように、①では事前にアプリ側でタグ部分をアンカーに置き換えて、修正したテキストを翻訳エンジンに送り、得られた翻訳結果の中のアンカー文字列をもとのタグ情報に戻せばよい。一方②では、原文と出力された訳文の間でそれぞれの単語がどのように対応しているかの情報を翻訳エンジンから取得できれば、原文に付与されていたタグを対応する訳文の単語につける処理が行える。例えば、統計翻訳ツールである Moses[149]においては、-t オプションを指定することにより訳文に単語対応情報が付与されて出力される。

```
> echo ’das ist ein kleines haus’ ¦ moses -f phrase-model/moses.ini -t > out
> cat out
this is ¦0-1¦ a ¦2-2¦ small ¦3-3¦ house ¦4-4¦
```

図５．１．１-６　Moses における単語対応出力 [149]

図５．１．１-６の例はドイツ語の das ist ein kleines haus が英語 this is a small house に翻訳された例であり、英語の this is が原文の０から１番目の語、すなわち das と ist に対応していることを示している。この情報を使えば、もしドイツ語の haus に何かタグがついていれば、英語の house にもそのタグをつければよいわけである。このように SMT にしろ RBMT にしろ原文と訳文の間の対応関係が翻訳エンジンから出力できれば、タグ復元が可能な翻訳アプリケーションを構築することができる。

(3) タグ復元を伴う翻訳処理の現状

本対象として調査した既存の機械翻訳ソフト、サービスは以下のとおりである。これらにおいては、全てのソフトとサービスにおいて、翻訳結果に原文の HTML のタグを復元できる HTML 翻訳機能があることが確認できた。ただし、ソフトによっては、基本的な事例においても本来復元すべき箇所にタグが付与されておらず原文文書におけるレイアウトが保たれない場合もあった。

149 Moses Statistical Machine Translation System User Manual and Code Guide, http://www.statmt.org/moses/manual/manual.pdf（最終検索日:2016 年 6 月 30 日）

表５．１．１　タグ復元に関して調査を行った機械翻訳ソフト並びにサービス一覧

The 翻訳プロフェッショナル v15 特許エディション[150]
j 北京特許エディション[151]
PAT-Transer V11[152]
ATLAS V14[153]
Google 翻訳[154]
Yahoo 翻訳[155]
エキサイト翻訳[156]
SYSTRAN[157]
Bing 翻訳[158]

（製品名若しくはサービス名は各社の商標若しくは登録商標である場合がある）

なお、上で述べたように、タグの復元機能は翻訳エンジンの外側でも実現できる機能であるため、特許文献向けの機械翻訳のサービスやアプリケーションを実現する際に、アプリケーションにタグ復元機能を持たせることも可能である。そのため、現時点で実用化されている機械翻訳エンジンやサービスの仕様や制限事項にとらわれる必要はなく、必要とするアプリケーションの機能仕様に応じてアプリ側で機能実装すればよいと考えられる。

5.1.2　タグを利用した翻訳

(1) 処理の概要

原文文書のタグを単純にそのまま訳文文書に復元する以外に、翻訳を実行する際にタグの情報をもとに以下のような制御をおこなって、翻訳の精度を改善できる可能性がある。

① 数式や化学式のように、翻訳しなくてもよい範囲を、タグ名に基づいて決定する。
② 「発明の名称」のようなタイトルや表の中の要素を名詞句として認識し翻訳するといった具合に、通常のテキストの翻訳とは異なる方式で翻訳する。
③ 人名や会社名、住所などは、翻訳しない、若しくは音訳するなど、それらの種別に

150 The 翻訳プロフェッショナル,http://pf.toshiba-sol.co.jp/prod/hon_yaku/premium_pat/index_j.htm（最終検索日:2016 年 6 月 30 日）
151 中国語特許翻訳, http://www.kodensha.jp/soft/jbpat/（最終検索日:2016 年 6 月 30 日）
152 PAT-Transer V12, http://www.crosslanguage.co.jp/products/pat-transer_v12/index.html（最終検索日:2016 年 6 月 30 日）
153 FUJITSU Software ATLAS, http://www.fujitsu.com/jp/products/software/applications/applications/atlas/（最終検索日:2016 年 6 月 30 日）
154 Google 翻訳, https://translate.google.co.jp/（最終検索日:2016 年 6 月 30 日）
155 Yahoo!翻訳, http://honyaku.yahoo.co.jp/url/（最終検索日:2016 年 6 月 30 日）
156 エキサイト 翻訳, http://www.excite.co.jp/world/（最終検索日:2016 年 6 月 30 日）
157 SYSTRAN, http://www.systransoft.com/（最終検索日:2016 年 6 月 30 日）
158 Bing 翻訳, https://www.bing.com/translator（最終検索日:2016 年 6 月 30 日）

応じたやり方で翻訳する。

(2) 現状

既存の機械翻訳ソフトにおいては、HTML 文書の翻訳機能において、翻訳が不要なタグを指定できるものがある。また、表を認識しその中の要素を通常のテキストとは異なる形式で翻訳できるソフトもある。いずれにしろ、上で述べたような機能が必要か否かはアプリケーションの機能として検討すべき課題であり、翻訳方式を SMT とするか RBMT とするかに関係なく、必要であればアプリケーション側で実装すればよく、翻訳エンジンが持つべき機能として考える必要はないと思われる。

5.2 オフィスソフトとの連携

マイクロソフト社が提供するオフィスソフト Microsoft Office では、ワープロソフト Word、表計算ソフト Excel、プレゼンテーションソフト PowerPoint の機能として翻訳機能が提供されており、メニューの中のコマンドを使用して編集中の文書若しくはユーザが指定した特定範囲の文字列に対する翻訳を行うことができる。そこで利用されているのは表５．１．１に掲載した Bing 翻訳である[158]。

表５．１．１に掲載した製品のうち、PC にインストールして使用するパッケージソフトウェアのほとんどでは、上と同様の翻訳機能を Microsoft Office の各種ソフトにマクロを組み込むことで利用できるようになっている[150][151][152][153]。また、Office のソフトを立ち上げなくても、ほとんどのソフトでは、ファイル翻訳機能の中で翻訳したいファイルを指定すれば、翻訳できるようになっている。

5.3 メールの翻訳

ここでは、5.2 で述べたオフィスソフトと同様、マイクロソフト社が提供するオフィスソフト Microsoft Office の一部として提供されているメールソフトである Outlook を例にとり、メール内容の翻訳機能について説明する。

Outlook には、Word などで提供されている翻訳機能と同様、メールソフトのメニューの中に翻訳機能が組み込まれており、ユーザが指定した範囲若しくはメール全体の翻訳が可能である。ただし、一般の文書と比較した場合にメールの翻訳で問題となるのは、引用の扱いである。メールを返信する際に、相手から受信したメールの一部を引用することがよくあるが、その場合、テキストの改行直後の非空白文字が引用を示す記号として用いられることが多い。具体例を図５．３に示す。

XX 様、

>こちらに到着する時刻をお知らせください。また宿
>泊するホテルは決まっていますか。

23 日の 17 時の予定です。まだホテルは予約していません。

図５．３　引用を含むメールの例

図５．３の例では行頭が ”> ” で始まる３行目、４行目が引用となっており、「宿泊」という語が途中で改行されてしまっている。これらを適切に扱うには、引用部分全体を認識したあと、翻訳時のテキストには含めるべきではない引用記号を削除し、なおかつ途中で改行が入っている部分を連結して本来の文の形に戻してから翻訳をしなければ適切な翻訳結果は得られない。

なお、引用記号は通常、メールの利用者が自由に設定できるため、常に ”> “ という文字列が使われるわけではない。そのため、引用記号をユーザが指定できるようになっていることが望ましい。ただこういったメールの翻訳機能自体は、5.1 で述べた構造化文書の翻訳機能と同様、通常のテキストの翻訳機能を提供する翻訳エンジンが利用できれば、メール翻訳のアプリ開発者が適宜実現できる機能であり、翻訳エンジン自体が持っていなければならない必須機能であると考える必要はない。

6． 言語別の状況

今回の調査対象の言語である①ベトナム語、②タイ語、③インドネシア語、④中国語、⑤韓国語、⑥英語、⑦日本語の７言語について、各言語を母語とする言語処理の専門家に調査を依頼し、特許翻訳及び言語処理のポイントとなる各国語の特徴をまとめた。また、調査対象の言語の言語資源と言語処理ツールについても示した（「2.言語リソース」及び「3.要素技術」も参照のこと）。

6.1 ベトナム語

6.1.1 ベトナム語の特徴

(1) 文字

ベトナム語で使用されている文字を図６．１．１に示す。

A,Ă, B,C, D,E, Ê, G, H, I, K, L, M, N, O, Ô, Ơ, P, Q, R, S, T, U, Ư, V, X, Y (a,ă,b,c,d,e,ê,g,h,i,k,l,m,n,o,ô,p,q,r,s,t,u,ư,v,x,y)

図６．１．１ ベトナム語で使用されている文字

(2) 分かち書き

ベトナム語の文は空白が用いられるが、それらは以下で述べる音節の区切りを示すものであって語の区切りを示すものではない。したがって日本語文に対する形態素解析と同様、単語の区切りを同定する処理が必要となる。

(3) 語の表記

ベトナム語において語を構成する基本単位である音節は５つの部分を持つ。すなわち、頭子音、介母音、主母音、末子音、声調記号である。例えば、音節 tuan（week）の場合、頭子音=t、介母音=u、主母音=â、末子音=n、声調記号=`（低アクセント記号）から構成される。しかし、これらのうち、主母音はいかなる音節においても必要であるが、それ以外の要素は使われない場合もある。例えば、音節 anh（brother）は頭子音、介母音、声調記号を持っていない。また、音節 hoa（flower）には末子音が存在しない。

ベトナム語の単語は１つ以上の音節を持つ。2 つ以上の音節を持つ語であっても英語などと違い音節の間に以下の例のように空白が入る。

例１ đại học （đại（大） học（学） = 大学）

また、以下の例のように、語の前後に接頭語や接尾語をつけて派生させて新たな語を作ることもできるが、この場合にも接頭語や接尾語の前後には空白が使用される。

例２　siêu thị （siêu-（超-） thị（市） → 超市=supermarket）
例３　ngôn ngữ học （ngôn ngữ（言語） -học（-学） → 言語学）

なお、古い表記法においては語を構成する音節を空白ではなくハイフンでつなぐ表記法もあったが現在は空白が用いられている。

また、語若しくは語の一部の音節を繰り返して構成される畳語というものがある[159]。

(4) 語順

ベトナム語は分析的言語（analytic language）若しくは孤立語（isolating language）に分類される言語であり、格、性、数、時制などの形態的指標を持たない SVO 言語である。また、名詞の修飾語は名詞の後ろに置かれる点は同じ SVO 言語に分類される英語とは異なる特徴である。このような言語の特徴に関する説明は wikipedia でも見つけることができる[160]。

6.1.2　ベトナム語処理のための言語リソース

ベトナム国内においては、Vietnam Language and Speech Processing（VLSP）National Project（KC01.01.05/06-10）と呼ばれるが国家プロジェクトが、ベトナム語の言語処理技術の開発のための最初の国家プロジェクトである[161]。このプロジェクトで構文タグ付きコーパスや辞書、対訳コーパスなどが開発されたが、実用に供するアプリケーションの開発のためには、質、量ともに不十分である。

(1) 対訳コーパス

・VLSP corpus

VLSP プロジェクトで開発された英越対訳コーパスで、対訳数は 10 万文対である [161]。

・EVBCorpus

ベトナムの情報技術大学がウイーン大学等と共同で開発した 80 万文対以上を含む英越対訳コーパスである[162]。これらのうち、4.5 万文対に対しては単語対応の情報がつけられている。本コーパスは関連ツールとともに Google Code にて公開されている[163]。

159 N. C. Mai, D. N. Vu, and T. P. Hoang. Foundations of Linguistics and Vietnamese. Education Publisher, 1997.
160 https://en.wikipedia.org/wiki/Vietnamese_language（最終検索日:2016 年 7 月 12 日）
161 VLSP project　http://vlsp.vietlp.org/（最終検索日:2016 年 7 月 12 日）
162 Quoc Hung Ngo, Werner Winiwarter, Bartholomaus Wloka,（2013）. "EVBCorpus - A Multi-Layer English-Vietnamese Bilingual Corpus for Studying Tasks in Comparative Linguistics", In Proceedings of the 11th Workshop on Asian Language Resources（ALR11）, IJCNLP2013 Workshop, pp. 1-9. AFNLP, 2013.
163 https://code.google.com/archive/p/evbcorpus/（最終検索日:2016 年 7 月 12 日）

・IWSLT コーパス

音声翻訳に関する評価型のワークショップである IWSLT 2015 の TED Talks に関するタスク用のリソースとして英越対訳コーパスが提供されている[164] [165]。

・English-Vietnamese parallel corpus（EVC）[166]

ベトナム国立大学（ホーチミン市）が開発した英越対訳コーパス。収録文数は約 25 万文対で、そのうち 17.5 万文対は英越対訳辞書の例文を利用したものである[167]。ライセンス条件は不明である。

(2) 単言語コーパス

ベトナム語の注釈付きコーパスが開発されている。

・Vietnamese Treebank

VLSP プロジェクトの中で開発された構文タグ付きコーパスである[161] [168] [169]。コーパスに採録されたテキストはベトナムの Youth Association が提供する Web のニュースサイト Thanh Nien 新聞[170]の記事から抽出された。タグ付与されている文の数は約１万で、現在の精度は約 90%である。Treebank の中には品詞タグ付きコーパスや単語分割コーパスも含まれている。

・総合研究大学院大学コーパス

日本の総合研究大学院大学、NII 並びにベトナムの社会科学・人文大学、情報科学大学により開発が進められているコーパスである[171]。約４万文のタグ付きデータが作成されており、精度は 90%以上である。現時点ではまだ公開されていない。

・Vietnamese Dependency Treebank（VnDT）

ベトナムのベトナム国立大学と日本の JAIST が、言語学的規則を使用して上で記載した

164 Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli and MarcelloFederico. Report on the 12th IWSLT Evaluation Campaign, IWSLT 2015

165 https://wit3.fbk.eu/mt.php?release=2015-01（最終検索日:2016 年 7 月 12 日）

166 http://www.clc.hcmus.edu.vn/?page_id=467&lang=en（最終検索日:2016 年 7 月 12 日）

167 Dinh Dien and Hoang Kiem: POS-Tagger for English-Vietnamese Bilingual Corpus. In Proc. of HLT-NAACL 2003 Workshop, WS4: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, pp.88-95, 2003.

168 Nguyen, P.-T., Vu, X. L., Nguyen, T. M. H., Nguyen, V. H., & Le, H. P.（2009）. Building a large syntactically-annotated corpus of Vietnamese. In Proceedings of LAW-3, ACL-IJCNLP

169 PT Nguyen, AC Le, TB Ho, VH Nguyen, Vietnamese treebank construction and entropy-based error detection - Language Resources and Evaluation, 2015.

170 http://thanhnien.vn/thoi-su/（最終検索日:2016 年 7 月 12 日）

171 Quy T. Nguyen, Y. Miyao, Ha T.T. Le, and Nga L.T. Nguyen, “Challenges and Solutions for Consistent Annotation of Vietnamese Treebank”, In Proceedings of LREC 2016.

Vietnamese Treebank を変換して作成した概念依存関係の treebank である[172]。タグ付されている文の数は約１万である。

(3) 対訳辞書

・Lacviet 社対訳辞書

Web 上の対訳辞書検索サービスとして Lacviet Computing 社が英越、越日などの辞書検索サービスを提供している[173]。データの規模やライセンス条件は不明である。

・VDict

ユーザ参加型で開発されている対訳辞書が Web で公開されている[174]。データの規模やライセンス条件は不明である。

(4) 単言語辞書

・VDict

対訳辞書でも記載した VDict において単語辞書も公開されている[174]。データの規模やライセンス条件は不明である。

(5) シソーラス、概念辞書

・Viet WordNet

英語の概念辞書として著名な WordNet のベトナム語版である[175]。３万語の synset を用いて５万語のデータが整備されている。商用利用はできない。

6.1.3 ベトナム語処理のための要素技術

(1) 形態素解析

ベトナム語の特徴の項で述べたように、文中の語の範囲を同定するのに、音節の分割に使われている空白は手掛かりとして使うことはできない。例えば、下の例４の文においては、２ないし３音節の要素ごとに分割する必要がある。

例４ Anh ay đi dạo ở công viên （He walks in the park）

↓

Anh_ay（He） đi_dạo （walk） ở công_viên（park）.

172 D. Q. Nguyen, D. Q. Nguyen, S. B. Pham, P. T. Nguyen, and M. L. Nguyen. From treebank conversion to automatic dependency parsing for Vietnamese. In Proceedings of the 19th International Conference on Application of Natural Language to Information Systems, NLDB'14, pages 196–207, 2014.

173 http://tratu.coviet.vn/（最終検索日:2016 年 7 月 12 日）

174 http://vdict.com/（最終検索日:2016 年 7 月 12 日）

175 http://viet.wordnet.vn/wnms/（最終検索日:2016 年 7 月 12 日）

・VLSP segmentation tool

VLSP プロジェクトで開発された分割ツールである。辞書と N-gram データ、正規表現に基づく処理により、closed data に対して 97.1%、open data（100 文）に対して 98.2%の精度が得られている。

・JVNSegmenter[176] [177]

ベトナムのベトナム国立大学と日本の JAIST で開発された segmentation tool である。タグ付きコーパスを用いて SVM（support vector machine）,CRF（Conditional Random Field）により学習を行っている。GNU GPLv2 でライセンスされており、商用利用可能である。

・vnTokenizer[178] [179]

ベトナム国立大学で開発されたベトナム語処理のツールキットである Vitk に含まれているトークナイザである。本処理器は大規模テキストを短時間で処理するため Apache Spark 上で並列分散処理できる。24 コアで RAM24GB のコンピュータ３台をクラスタで用いることで 200 万音節のテキストを 20 秒で分割処理できる。その精度は 97%である。また、品詞タグ付けについては、上記の性能のマシン１台で毎秒 110 万トークンの処理ができ、精度は 95%である。GNU GPL でライセンスされており商用利用可能である。

・VLPS part of speech tagging tool

VLSP プロジェクトで開発された品詞タガー。MEM（最大エントロピーモデル）と CRF を用いている。Vietnamese TreeBank に収録されている品詞タグのついた２万文のデータとベトナム語辞書を用いて訓練を行っている。品詞タグの種類は 18 で、精度は 92%であった。

・RDRPosTagger[180] [181]

transformation ベースの学習法を用いた品詞タガーで、処理精度は 92.6%である。文献

176 Cam Tu Nguyen, Trung Kien Nguyen, Xuan Hieu Phan, Le Minh Nguyen and Quang Thuy Ha　Vietnamese Word Segmentation with CRFs and SVMs: An Investigation, The 20th Pacific Asia Conference on Language, Information, and Computation（PACLIC）, 1st-3rd November, 2006, Wuhan, China

177 http://jvnsegmenter.sourceforge.net/（最終検索日:2016 年 7 月 12 日）

178 Phuong, L. H., Huyen, N. T. M., Azim, R., & Vinh, H. T.（2008）. A hybrid approach to word segmentation of Vietnamese texts. In Proceedings of the 2nd international conference on language and automata theory and applications. Springer LNCS 5196, Tarragona, Spain, 2008

179 https://github.com/phuonglh/vn.vitk（最終検索日:2016 年 7 月 12 日）

180 Dat Quoc Nguyen, Dai Quoc Nguyen, Dang Duc Pham and Son Bao Pham. RDRPOSTagger: A Ripple Down Rules-based Part-Of-Speech Tagger. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics（EACL）, pp. 17-20, 2014.

181 http://rdrpostagger.sourceforge.net/（最終検索日:2016 年 7 月 12 日）

[182]には、このツールと比較し、さらに高い性能（約 94%）のツールについての記載がある。そのツールでは大規模な平文コーパスで訓練した word cluster model を用いている。

なお、形態素解析の一要素技術として固有表現抽出（Named Entity Recognition）がある。94%の処理精度が得られた研究報告[183] [184]もあるが、そこで用いられたデータは現在入手できる状態ではない。

(2) 構文解析、係り受け解析

・VLSP chunking tool[185]

VLSP プロジェクトの中で開発された CRF とオンライン学習を用いたチャンキングツール（チャンカ）。訓練に用いられたデータは 9,000 文の品詞タグ付きコーパスで、94%の処理精度が報告されている。

・vTools[186] [187]

最大エントロピーモデルを用いたツールである。性能、ライセンス条件などは不明である。

・VnDP: A Vietnamese dependency parsing toolkit[170] [188]

VLSP プロジェクトで開発された Treebank を変換して作成した dependency treebank を用いた係り受け解析器である。訓練データが小さいため精度は約 77%と、それほど高くない。学術目的の利用しかできない。

・Vitk -- A Vietnamese Text Processing Toolkit[177]

ベトナム国立大学で開発されたベトナム語処理のツールキットである Vitk に係り受け解析器が含まれている。公開されているツールの処理精度は不明であるが、同ツールの開発者による研究発表では係り受け解析の精度として 73.21%と記載されているものがある

182 Le Minh Nguyen, Xuan Bach Ngo, Viet Cuong Nguyen, Quang Nhat Minh Pham and Akira Shimazu, A Semi-Supervised Learning Method for Vietnamese Part-of-Speech Tagging, In Proceedings KSE 2010.

183 Pham Thi Xuan Thao, Tran Quoc Tri, Dinh Dien, Nigel Collier, Named Entity Recognition in Vietnamese using classifier voting, ACM Transactions on Asian Language Information Processing（TALIP） 6（4）, 2007.

184 Nguyen, D.B., Hoang, S.H., Pham, S.B., Nguyen, T.P., 2010. Named Entity Recognition for Vietnamese, in: Proc. of the Second international conference on Intelligent information and database systems: Part II, Springer-Verlag, Berlin, Heidelberg. pp. 205–214.

185 Le Minh Nguyen, Huong Thao Nguyen and Phuong Thai Nguyen, An Empirical Study of Vietnamese Noun Phrase Chunking with Discriminative Sequence Models. In Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pages 9–16, Suntec, Singapore, 6-7 August 2009. c 2009 ACL and AFNLP

186 Mai-Vu Tran, Duc-Trong Le（2013）. vTools: Chunker and Part-of-Speech tools, RIVF- VLSP 2013 Workshop

187 https://github.com/lupanh/vTools（最終検索日:2016 年 7 月 12 日）

188 http://vndp.sourceforge.net/（最終検索日:2016 年 7 月 12 日）

[189]。GNU GPL でライセンスされており商用利用可能である。

・Dalat 大学の係り受け解析器[190]

ベトナムの Dalat 大学と国立大学で開発された係り受け解析器である。Vietnamese Treebank を変換して作成した概念依存関係の treebank を用いて学習させた。解析器自体は外部の解析ツールである MaltParser を用いている。現時点でツールとして公開されているわけではない。

・ベトナム国立大学の統計的構文解析器[191]

VLSP プロジェクトの中で開発された統計的な構文解析器である。語彙化した確率文脈自由文法を Vietnamese Treebank から学習した。精度は F 値で 78%である。研究報告だけであり、ツールとして公開されているわけではない。

6.1.4 機械翻訳

ベトナム語の機械翻訳に関する研究は英語との間の翻訳を対象として行われているものがほとんどである。処理精度は、Google による翻訳サービスを凌駕するレベルには達しているとはいえないのが現状であるが、以下に代表的な事例を挙げる。

・Lacviet 社の機械翻訳サービス[192]

対訳辞書の項で取り上げた Lacviet 社が提供している英越機械翻訳サービスである。英語からベトナム語への翻訳のみで、ベトナム語から英語への翻訳はできない。翻訳方式はトランスファ方式の RBMT である。

・EVTRAN

ベトナム・国立技術推進センター（National Center for Technological Progress：Nacentech）が開発した英越 RBMT。1999 年に商用化され、2005 年にリリースされた Ver.3.0 では語彙数が 53 万となっている[193]。現在は単語辞書の項で紹介した VDict のサイト[194]で提供されている翻訳サービスに使われている。

---

189 Le-Hong, P., A. Roussanaly, and T M H. Nguyen, "A Syntactic Component for Vietnamese Language Processing", Journal of Language Modelling, vol. 3, issue 1, pp. 145-184, 2015.

190 T.-L. Nguyen, V.-H. Nguyen, T.-M.-H. Nguyen, and P. Le-Hong. Building a treebank for Vietnamese dependency parsing. In Proceedings of RIVF, pages 147–151. IEEE, 2013.

191 AC Le, PT Nguyen, HT Vuong, MT Pham, TB Ho An experimental study on lexicalized statistical parsing for Vietnamese, In Proceedings KSE 2009.

192 Lac Viet Translation: http://tratu.coviet.vn/hoc-tieng-anh/dich-van-ban.html（最終検索日:2016 年 7 月 12 日）

193 http://www.jaist.ac.jp/~bao/talks/MachineTranslationinVN.pdf（最終検索日:2016 年 7 月 12 日）

194 http://vdict.com/?autotranslation#translation（最終検索日:2016 年 7 月 12 日）

・UNS-VNUHCM の機械翻訳システム[195] [196]

UNS-VNUHCM（Univ. of Natural Sciences, Vietnam National Univ. in HCMC）が研究している機械翻訳システムである。タグ付きの対訳コーパスから自動で抽出した変換規則（語彙規則並びに構造変換規則）を用いて翻訳を行うのが特徴である。2003 年以降研究成果は発表されていないため、研究は停止していると考えられる。

・EVSMT[197]

JAIST で研究が行われている SMT システムである。VLSP プロジェクトで開発した言語資源を活用して開発が行われたが、2009 年の研究報告以降、実用化に関する発表はなされていない。

6.2 タイ語

6.2.1 タイ語の特徴

ここでは、機械翻訳を始めとする各種言語処理の対象として見た場合のタイ語の特徴を簡単に記述する[198] [199] [200]。

(1) 文字

主にタイ語の表記に使われているタイ文字は表音文字である。子音は以下の 42 文字あり、この他に廃字になった２文字がある。母音は、基本母音が９種類、二重母音と複合母音がそれぞれ３種類ずつあり、基本母音９種類にはそれぞれ長母音と単母音がある。一つの母音を表すのに複数の母音文字を組み合わせて表記されることがあるため、unicode においては 15 文字が母音用の文字として定義されている[201]。タイ語の子音と母音を図６．２．１に示す。

子音：ก、ข、ค、ฆ、ง、จ、ฉ、ช、ซ、ฌ、ญ、ฎ、ฏ、ฐ、ฑ、ฒ、ณ、ด、ต、ถ、ท、 ธ、น、บ、ป、ผ、ฝ、พ、ฟ、ภ、ม、ย、ร、ล、ว、ศ、ษ、ส、ห、ฬ、อ、ฮ
母音：ะ、ั、า、ำ、ิ、ี、ึ、ื、ุ、ู、ฺ、เ、แ、โ、ใ、ไ

図６．２．１　タイ語の子音と母音

195 Dinh, D., Hoang, K., & Eduard, H.（2003）. BTL: A hybrid model in the English-Vietnamese machine translation system. Proceedings of the Machine Translation Summit IX.

196 Dinh, D., Thuy, N., Xuan, Q., & Chi, N.（2003）. A hybrid approach to word-order transfer in the E-V machine translation system. Proceedings of the Machine Translation Summit IX.

197 Ho, T. B., Pham, N., Ha, T., & Nguyen, T.（2009）. Issues and first phase development of the English-Vietnamese translation system EVSMT1.0. Proceedings of the third Hanoi Forum on Information — Communication Technology

198 Sornlertlamvanich V. and Pantachat W. “Information-based Language Analysis for Thai”, ASEAN Journal on Science & Technology for Development, Malaysia. Vol. 10 No. 2, pp. 181-196, 1993.

199 Muraki, K., Sornlertlamvanich, V., et al.（1989） “Thai Dictionary for Multi-lingual Machine Translation System”, Computer Processing of Asian Languages（CPAL）. AIT, 211-220.

200 Panupong, V.（1984） “The Structure of Thai Grammatical System”, Ramkhamhaeng Univ. Press.

201 タイ語文字コード表, http://www.unicode.org/charts/PDF/U0E00.pdf（最終検索日:2016 年 7 月 4 日）

(2) 分かち書き

下の例１に示すとおり、タイ語のテキストは英語のような空白による分かち書きは行われない。

例１：นักเรียนถูกครูทำโทษ

/nakrian/thuuk/khruu/longthoot/

"The student is punished by the teacher."

日本語の句点のような文の境界を示す記号もない。ただし、空白が全く使われないというわけではなく、文やパラグラフを読みやすくしたり、曖昧性を減らすために使われたりすることもある。いずれにしろ、空白や文末記号の存在を前提とすることはできないため、タイ語の処理においては、パラグラフから文を切り出す処理や、文を単語に分割する処理が不可欠となる。

単語分割処理は現在においてもタイ語の形態素解析処理の重要な課題である。他の言語同様、タイ語においても接辞を基本となる語に付与することで新しい語が作られる。そのため、辞書に単語で登録するか複合語を登録するかが問題となる。例えば、“การแปลภาษาด้วยคอมพิวเตอร์”（ /kaan/plxx/phaasaa/duai/khoomphiuter/ ）は５つの単語（"/kaan/", "/plxx/", "/phaasaa/", "/duai/", "/khoomphiuter/"）から構成されているが、分割の仕方には４つの仕方がある。

例２　การแปลภาษาด้วยคอมพิวเตอร์

/kaan/plxx/phaasaa/duai/khoomphiuter/

“Translation with computer” or “Machine Translation”

５つに分ける場合： /kaan/, /plxx/, /phaasaa/, /duai/, /khoomphiuter/

４つに分ける場合： /kaan/plxx/, /phaasaa/, /duai/, /khoomphiuter/

３つに分ける場合： /kaan/plxx/phaasaa/, /duai/, /khoomphiuter/

全く分割しない場合： /kaan/plxx/phaasaa/duai/khoomphiuter/

(3) 多義語の問題

ほとんどのタイ語の単語は複数の意味を持つ。語の使用頻度が高いほど、より多くの派生した意味を持つ。これは言葉をなるべく簡単に使いたいことからくる自然の要求である。しかし、その一方でそれらの違いを区別するための制約条件も必要になる。その制約条件には文法的な役割を示す品詞や動詞パターン、語用論的な隣接語の情報などが含まれる。例えば、例３の caak という語の場合、少なくとも３つの意味がある。

例３　จาก

/caak/　#POS.{V}, #CP.{LEAVE}
#POS.{PREP}, #CP.{FROM}
#POS.{N}, #CP.{NIPA^PALM}

(4) 屈折

タイ語は孤立語であり単音節言語である。英語において動詞の原形を過去形や現在分詞に変化させるような屈折現象はない。そのため、時制や態、アスペクト、モダリティを表す語が動詞の前後に置かれる。例４は受け身の例である。受け身の機能を持つ thuuk という語が使われている。

例４　นักเรียนถูกครูทำโทษ

/nakrian/　/thuuk/　/khruu/　/longthoot/
student　passive marker　teacher　punish
"The student is punished by the teacher."

時制も同様である。まだ起こっていない事象を示す“จะ”（/ca/ = will）という語がタイ語唯一の時制マーカーである。この語が用いられていなければ既に行った事象とみなすことができるが、現在の状態なのか、過去に起きたことなのかは文脈から読み取る必要がある。

(5) 語順

タイ語の語順は SVO であって、英語と同様である。ただし、修飾は後置修飾で、修飾句は修飾される語の後ろに置かれるという点が英語とは異なる。この語順は後述するインドネシア語でも同じである。

例５　ผมเล่นกีต้าร์　（私はギターを弾く）
ผม（私）　เล่น（弾く）　กีต้าร์（ギター）

例６　อาหารเผ็ด　（辛い料理）
อาหาร（料理）　เผ็ด（辛い）

6.2.2　タイ語処理のための言語リソース

タイ語の言語リソースのほとんどは公開されているものはほとんどなく、研究機関若しくは研究者が自らの研究目的で開発し、公開せずに使用している場合が多い。したがって

商用利用を前提としたライセンス条件などは開発者と個別に交渉する必要がある。

(1) タイ語と外国語間の対訳コーパス

特許以外の主な対訳コーパスとして以下がある。

・ASEAN MT project の対訳コーパス

ASEAN MT プロジェクトにてタイを含めた 10 か国語の旅行会話の対訳コーパスが収集された[202]。ただし、コーパスの規模はいずれも 2 万文と比較的小規模なものにとどまっている。これらのコーパスをもとに開発された SMT のサービスが Web で公開されている[203]。

(2) タイ語平文コーパス／注釈付きコーパス

・ORCHID コーパス[204] [205] [206] [207]

ORCHID コーパスとは、日本の郵政省通信総合研究所（CRL、現在は NICT）とタイの National Electronics and Computer Technology Center（NECTEC）の共同研究で開発された品詞タグ付きコーパスである。コーパスのサイズは約 40 万語で、NECTEC の年次会議の予稿集をもとに作成された。公開に関する情報はない。

・NECTEC word annotated corpus

NECTEC word annotated corpus は NECTEC が主催した BEST（Benchmark for Enhancing the Standard for Thai Language Processing）コンテストのために開発した 500 万語の単語タグ付きコーパスである。このコーパスの開発により、タイ語の単語分割の精度が若干向上した。SNLP 2016[208]では、タイ語の単語分割に関する 2 つの論文が発表されている。一つは辞書ベースの手法に条件付き確率場（conditional random field, CRF）を用いることで精度の改善を図ったものである。もう一つは単語分割にドメイン適応の手法を取り入れることを試みたものである。現時点では本コーパスに関する公開の情報は見当たらないため、研究目的限定でタイの研究者コミュニティ内でのみの利用しかできないものと考

202 2015 年 5 月 15 日時点でのプロジェクトの状況, http://www.aseanmt.org/index.php?q=index/status_update（最終検索日:2016 年 7 月 5 日）

203 ASEAN Machine Translation, http://www.aseanmt.org/mt/（最終検索日:2016 年 7 月 5 日）

204 Thatsanee Charoenporn, Virach Sornlertlamvanich and Hitoshi Isahara. Building A Large Thai Text Corpus---Part-Of-Speech Tagged Corpus: ORCHID---.Proceedings of the Natural Language Processing Pacific Rim Symposium, 1997.

205 Virach Sornlertlamvanich, Thatsanee Charoenporn and Hitoshi Isahara. ORCHID: Thai Part-Of-Speech Tagged Corpus. Technical Report Orchid TR-NECTEC-1997-001, National Electronics and Computer Technology Center, Thailand, pp. 5-19, Dec 1997.

206 Virach Sornlertlamvanich, Naoto Takahashi and Hitoshi Isahara. Thai Part-Of-Speech Tagged Corpus: ORCHID. Proceedings of Oriental COCOSDA Workshop, pp 131-138, 1998.

207 Virach Sornlertlamvanich, Naoto Takahashi and Hitoshi Isahara. Building a Thai Part-Of-Speech Tagged Corpus（ORCHID）. The Journal of the Acoustical Society of Japan（E）, Vol.20, No.3, pp 189-140, May 1999.

208 SNLP2016 The Eleventh International Symposium on Natural Language Processing, http://www.snlp2016.net/（最終検索日:2016 年 7 月 5 日）

えられる。

(3) 対訳辞書

・LEXiTRON[209]

コーパスベースの初めての辞書である。第１版はタイ語 1.1 万語と英語 9,000 語を収録しており、各タイ語には品詞と動詞パターンの情報の他、同義語、反義語、例文、語のグループ、対応する英語が付与されている。それぞれのエントリーはインターネット上のタイ語のページにおける出現頻度が確認されている。最新の更新は 2009 年で第３版がリリースされている。オープンソースとして提供されており任意の目的での利用が可能である。

・CICC タイ語基本語辞書[210]

（財）国際情報化協力センター（CICC）が 1995 年に作成した辞書の一つに５万語のタイ語基本辞書が含まれている。辞書のエントリーには対応する英語も付与されているためタイ英対訳辞書と見なすことができる。また、コンピュータ、電気関係の専門用語を収録した 2.7 万語の辞書もある[211]。言語資源協会（GSK）から配布されているが、商用利用はできない。

(4) タイ語単言語辞書

公開されている大規模なタイ語単語辞書の情報はない。当然のことながら、対訳辞書のデータのタイ語の情報を利用することは可能である。

(5) シソーラス、概念辞書

・Asian WordNet[212] [213] [214]

米プリンストン大学で開発された英語の概念辞書と対訳辞書を用いて半自動で構築され

209 Thai-English Electronic Dictionary LEXiTRON, http://lexitron.nectec.or.th/2009_1/index_en.php?q=index（最終検索日:2016 年 7 月 5 日）

210 GSK2006-A-4 CICC タイ語基本語辞書, http://www.gsk.or.jp/catalog/gsk2006-a-4/（最終検索日:2016 年 7 月 5 日）

211 GSK2006-A-5 CICC 専門語辞書, http://www.gsk.or.jp/catalog/gsk2006-a-5/（最終検索日:2016 年 7 月 5 日）

212 Virach Sornlertlamvanich, Thatsanee Charoenporn, Hitoshi Isahara. Language Resource Management System for Asian WordNet Collaboration and Its Web Service Application, Proceedings of The Seventh International Conference on Language Resources and Evaluation（LREC）, Mediterranean Conference Center（MCC）, Malta, May 17-23, 2010.

213 Virach Sornlertlamvanich, Thatsanee Charoenporn, Kergrit Robkop, Chumpol Mokarat, and Hitoshi Isahara. Review on Development of Asian WordNet, JAPIO 2009 Year Book, Japan Patent Information Organization, Tokyo, Japan, 2009.

214 Sareewan Thoongsup, Thatsanee Charoenporn, Kergrit Robkop, Tan Sinthurahat, Chumpol Mokarat, Virach Sornlertlamvanich and Hitoshi Isahara. Thai Wordnet Construction, Proceedings of The 7th Workshop on Asian Language Resources（ALR7）, Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics（ACL） and the 4th International Joint Conference on Natural Language Processing（IJCNLP）, Suntec, Singapore, August 6-7, 2009.

た概念辞書である。アジアでは 13 言語において同様の辞書が作られているが、タイ語については約 7.2 万概念が定義されている。

6.2.3　タイ語処理のための要素技術

(1) 形態素解析

タイ語の単語分割[215]については、いくつかの研究が論文として報告されている。最長一致（longest matching）に基づく手法や最大一致（maximal matching）に基づく手法など辞書ベースの手法を初めとして、品詞トライグラム（3-gram）によるタグ付けのような統計的モデルや CRF やフィーチャベースの機械学習モデルを導入することで改善が図られている。

・SWATH[216] [217]

SWATH はオープンソースとして公開されているタイ語の単語分割器で、最長一致、最大一致、品詞バイグラム（2-gram）の３つのアルゴリズムを選択できる。

上記以外にもオープンソースの解析器が公開されているが、理論的な背景が提供されているものはほとんどない。

(2) 固有表現抽出と意味関係抽出

固有表現抽出（Named Entity Recognition, NER）と意味関係抽出（Semantic Relation Extraction,SRE）はテキスト間の関係を理解するなどテキストの本質的な情報を表現するために必須の要素である。近年多くの研究分野において、深層学習を初めとする機械学習のアプローチにより性能を改善する可能性が示されている。タイ語処理の意味関係抽出の研究においても、その成果が報告されている[218] [219] [220]。ただし、ツールとして整備され公開されているようなものは見当たらない。

---

215 Virach Sornlertlamvanich. Word Segmentation for Thai in Machine Translation System. Machine Translation, National Electronics and Computer Technology Center, Bangkok. pp. 50-56, 1993.（in Thai）

216 Surapant Meknavin, Paisarn Charoenpornsawat, and Boonserm Kijsirikul, 1997. Feature-based Thai Word Segmentation. In Proceedings of the Natural Language Processing Pacific Rim Symposium 1997（NLPRS'97）, Phuket, Thailand.

217 Software: SWATH - Thai Word Segmentation, http://www.cs.cmu.edu/~paisarn/software.html（最終検索日:2016 年 7 月 5 日）

218 Virach Sornlertlamvanich and Canasai Kruengkrai. Effectiveness of Keyword and Semantic Relation Extraction for Knowledge Map Generation, Proceedings of The Second International Workshop on Worldwide Language Service Infrastructure（WLSI）, Kyoto University, Kyoto, Japan, January 22-23, 2015.

219 Watchira Buranasing, Virach Sornlertlamvanich, and Thatsanee Chalernporn. Semantic Relation Extraction for Extensive Service of a Cultural Database, Proceedings of The Tenth Symposium on Natural Language Processing（SNLP）, Phuket, Thailand, October 28-30, 2013.

220 Canasai Kruengkrai, Virach Sornlertlamvanich, Watchira Buranasing, and Thatsanee Charoenporn. Semantic Relation Extraction from a Cultural Database, Proceedings of Workshop on South and Southeast Asian NLP, COLING2012, Mumbai, India, December 8-15, 2012.

(3) 機械翻訳

機械翻訳はタイでの自然言語処理研究において長い歴史を持つ研究テーマの一つである[221] [222]。現在タイにおいて注目されている研究の流れには以下の２つがある。一つは中国語とタイ語の間の階層型フレーズベースの統計翻訳であり、もう一つは ASEAN MT プロジェクトに参加している研究機関と連携にもとづく研究である。

・ParSit[223] [224]

ParSit は、Web ベースの最初の英泰機械翻訳サービスである。タイ語の生成部の開発には CICC による多言語機械翻訳システムでのタイ語処理の成果が用いられている。サービスは現在を行われているが、サービスサイトは調査時点で準備中となっており、サービス提供の実態は確認できていない。また、ライセンス情報なども不明である。

(4) 書記素音素変換、トランスリテレーション等

書記素音素（Grapheme-to-phoneme,GTP）変換[225] [226]はテキスト形式のタイ語を発音形式に変換する処理であり、例えば、タイ語“ภาษาไทย”を/phaa-saa-thai/といった形に変換する。この処理は主に音声合成（text-to-speech）で必要とされる技術である。これと類似しているが目的が異なる技術にトランスリテレーション（字訳、音訳）[227]技術がある。この技術はタイ語の単語を音声上等価若しくは類似したアルファベット表現に変換するのに用いられ、主にタイ語を他の言語に翻訳する際に何らかに理由でタイ語をローマナイズする必要があるような場合に用いられる。

・Soundex

Soundex という名称は、各言語において元のスペルを一般化された発音形式の表現に変換するための機能の名称として広く使われているものであり、音の類似した単語を検索するために使用されるためデータベースなどの処理系で提供されている。また、スペルチェックに用いられることもある。タイ語についても、このような機能を提供するソフトが開

---

221 Virach Sornlertlamvanich. Another Decade of Thai Language Processing Research. International Symposium on Multilingual Machine Translation（MMT'94）, Tokyo, Japan. pp. 56-60, 1994.

222 Virach Sornlertlamvanich. MT Research in Thailand and Linguistics and Knowledge Science Laboratory（LINKS）. AAMT, 1994.

223 Virach Sornlertlamvanich, Paisarn Charoenpornsawat, Monthika Boriboon and Lalida Boonmana. ParSit: English-Thai Machine Translation Services on Internet. 12th Annual Conference, ECIT and New Economy, National Electronics and Computer Technology Center, Bangkok, pp. 427-482, June 2000.（in Thai）

224 http://www.suparsit.com/（公開準備中）

225 Pongthai Tarsaku and Virach Sornlertlamvanich. Grapheme to Phoneme for Thai, Proceedings of NLPRS, NII, Japan, Nov 2001.

226 Pongthai Tarsaku, Virach Sornlertlamvanich and Rachod Thongpresirt. Thai Grapheme-to-Phoneme Using Probabilistic GLR Parser, Proceedings of Eurospeech 2001, Aalborg, Denmark, Sept 2001.

227 Thatsanee Charoenporn, Ananlada Chotimongkol and Virach Sornlertlamvanich. Automatic Romanization for Thai. Proceedings of the Second International Workshop on East-Asian Language Resources and Evaluation（ORIENTAL COCOSDA '99）, pp 137-140, 1999.

発されている[228] [229]。

6.3 インドネシア語

6.3.1 インドネシア語の特徴

(1) 文字

例１に示すように、インドネシア語はローマ字のアルファベットを用いて表記される。

例１　Saya membeli sebuah buku.　（Saya:I、membeli:buy、sebuah:a、buku:book）

使用される文字は英語と同じで、以下に示す 26 文字である。

A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z

(2) 分かち書き

例１から分かるようにインドネシア語の文は空白を用いて分かち書きされる。したがって英語と同様、インドネシア語の文を単語に分割する場合には、これらの空白をもとに行えばよい。

(3) 語形変化

インドネシア語は接辞によって派生する。接辞には次のようなものがある。

①　命令や疑問を示す接尾辞 -kah、-lah、-tah、-pun

動詞や形容詞の語根に付き、疑問や命令などの意味を付加する。

例２　beli （buy） + -lah → belilah （please buy）

②　所有の意味を付与する接尾辞 -ku、-mu、-nya

以下の例のように、語根につき、元の名詞に対する所有の意味を付加する。

例３　buku （book） + -ku （my） → bukuku （my book）

なお、②は①の接尾辞と組み合わせて使用することも可能である。その場合、①の接尾辞は②の接尾辞の後に付加される。

例４　buku （book） + -ku （my） + -lah → bukukulah （of course my book）

（注：接尾辞 lah は、例２と例４に示すように、文脈によって意味が変わる。）

[228] PPA Soundex 2, http://www.puttipan.com/soundex/index_en.html（最終検索日:2016 年 7 月 5 日）

[229] A Thai Soundex System for Spelling Correction, https://linux.thai.net/~thep/soundex/soundex.html（最終検索日:2016 年 7 月 5 日）

③　接頭辞　ber-、di-、ke-、me-、pe-、se-、te-

これらの接頭辞は、語根の先頭の文字によって変化する。例えば、me-は mem-、men-、meng-、meny-のようになることもある。これらの接頭辞は、一つの語根に最大３つまで付加することができる。

例５　se- + pe- + ke- + tahu （know） + -an + -ku → sepengetahuanku（as far as I know）

例５では、“ tahu ”（know）という語に対し、se-、peng-（pe-の異形)、ke-の３つの接頭辞と-an、-ku という２つの接尾辞がついて、一つの語となっている。

④　他動詞化、名詞化を行う接尾辞　–i、-kan、-an

これらの接尾辞は、語根に対し、一つしか付与することができない。

例６　beli （buy） + -kan → belikan （inquiry to buy）

⑤　共接辞　ber--an、ke--an、me--i、me--kan、di--i、di--kan

接頭辞と接尾辞が同時にある語幹に付与される場合、それらのペアを共接辞と呼ぶ。例えば、下の例７においては、買うという意味の beli という語の前後に mem-と-kan という接辞がついて、与える、供給するという意味の membelikan という動詞が派生されている。

例７　mem- + beli （buy） + -kan → membelikan （do a favor to buy）

インドネシア語で最もよく使われる共接辞は ber--an と ke--an である。ber--an は動作や状態の反復や継続（来る->次々とやって来る)、相互関係（愛->愛し合う）などの意味の語を派生させる。また、ke--an は被害の受け身（雨->雨に降られる)、自発（見る->見える）などの意味の語を派生させる。

(4) 語順

例１に示した例文から分かるようにインドネシア語の語順は SVO であって、英語と同様である。修飾については、以下に示すように基本は後置修飾である。

①　形容詞による名詞の修飾

英語では連体修飾する形容詞は名詞の前に置かれるが、インドネシア語では形容詞は

名詞の後ろに置かれる。

例８　e（red:adj） + e（car:n） → e（red car）
　　i（merah:adj） + i（mobil:n） → i（mobil merah）

例８において、e（...）は英語の語を、i（…）はインドネシア語であることを示す。また、カッコ内のコロンの右側に記載された n, adj はそれぞれ品詞が名詞、形容詞であることを表している。したがって i（merah:adj）は、インドネシア語の merah という形容詞を、i（mobil:n）は mobil という名詞を示しており、それぞれ英語の red と car に対応する。しかし、これらの２語から構成される英語の ”red car ” に対応するインドネシア語は形容詞が後置されて ”mobil merah ” となっている。

②　名詞修飾

２語の名詞により複合語を構成する場合、ヘッドとなる名詞を修飾する名詞は英語では前置されるが、インドネシア語では形容詞の場合と同様、ヘッドの後ろに置かれる。

例９　e（intelligence:nm） + e（system:n） → e（intelligence system）
　　i（kecerdasan:nm） + i（sistem:n） → i（sistem kecerdasan）

上記において名詞修飾語には nm という品詞が付与されている。インドネシア語の“ kecerdasan ”（intelligence）は“ sistem ”（system）の後ろに置かれている。

③　冠詞や指示詞、数量詞等の語順

上で述べたように、インドネシア語では名詞を修飾する形容詞や名詞修飾語は修飾される名詞の後ろに置かれると述べたが、不定冠詞（例 10）は英語と同様、名詞の前に置かれる。一方、定冠詞（例 11）と指示詞（例 12）は形容詞などと同様に後置される。

例 10　e（a:ar） + e（network:n） → e（a network）
　　i（sebuah:ar） + i（jaringan:n） → i（sebuah jaringan）

例 11　e（the:ar） + e（network:n） → e（the network）
　　i（tersebut:ar） + i（jaringan:n） → i（jaringan tersebut）

例 12　e（these:d） + e（networks:n） → e（this networks）
　　i（ini:d） + i（jaringan-jaringan:n） → i（jaringan-jaringan ini）

英語の few や many、little などにあたる数量詞は前置される。

例 13　e（few:q） + e（cars:n） → e（few car）
　　i（sedikit:q） + i（mobil-mobil:n） → i（sedikit mobil）

注：上記例文において複数の車を示すインドネシア語は mobil-mobil になっている。これはインドネシア語においては名詞をハイフンで連続することで複数を示すことができるためである。しかし、“ sedikit ”（few）を付けることで数量の意味が明確化されるため、名詞の連続による複数形を構成する必要がなくなり、 ”sedikit mobil ”となっている。

また、数詞の扱いも上で述べた数量詞と同様である。

例 14　e（two:num） + e（cars:n） → e（two cars）
　　i（dua:num） + i（mobil-mobil:n） → i（dua mobil）

④　形容詞を修飾する副詞

形容詞を修飾する副詞は形容詞の前に置かれる。そのため、名詞を修飾する形容詞に対して副詞で修飾する場合、例 15 に示すように、副詞は名詞と形容詞の間に挟まる形となる。

例 15　e（very:adv） + e（big:adj） + e（car:n） → e（very big car）
　　i（sangat:adv） + i（besar:adj） + i（mobil:n） → i（mobil sangat besar）

(5) 借用語

インドネシア語には、主にサンスクリット語、アラビア語、ポルトガル語、オランダ語、英語などから語が取り入れられている。語の取り入れにあたっては、インドネシア教育省（Ministry of Education of Indonesia）によるチェックが行われている。

語を取り入れる場合のやり方には次の２つがある。

①　Adoption

元の言語のスペル、発音を変更せずにそのまま取り入れることを adoption と呼ぶ。これらの語には、hot dog, reshuffle, shuttle cock, plaza, supermarket などがある。

② Adaptation

インドネシア語の表記の体系に合うようにスペルを修正して取り入れることを adaptation と呼ぶ。adaptation の事例には次のようなものがある。

・オランダ語の aa は a に変更される

例 16 paal → pal
baal → bal
octaaf → oktaf

・母音 a,u,o 並びにすべての子音の前の c は k に変更される

例 17 calomel → kalomel
construction → konstruksi
cubic → kubic
coup → kup
classification → klasifikasi
crystal → kristal

6.3.2 インドネシア語処理のための言語リソース

(1) 対訳コーパス

特許以外の主な対訳コーパスとして以下がある。

・IDENTIC Corpus[230]

チェコのプラハ・カレル大学で開発された約 100 万語（4.5 万文）からなるコーパスである。種々のデータソースから収集された異なるジャンルのテキストが収録されている。品質を高めるために、前処理としてアラインメントやスペルミスの修正を人手で行っている。データは平文形式の他に形態素解析結果の情報を付与した ’morphologically enriched’ な形式のデータも作成されている。コーパスに関する情報は IDENTIC のホームページで入手できる。

・PANL-BPPT Parallel Corpus[231]

インドネシアの BPPT（Agency for the Assessment and Application of Technology:インドネシア技術応用評価庁）によって開発されたコーパス。Penn Treebank Corpus から抽

230 S. D. Larasati, “IDENTIC Corpus: Morphologically Enriched Indonesian-English Parallel Corpus.,” in LREC, 2012, pp. 902–906.

231 PAN Localization Phase II Outputs, http://panl10n.net/english/OutputsIndonesia2.htm（最終検索日:2016 年 6 月 30 日）

出した 50 万語の英文を人手でインドネシア語に翻訳した。さらにインドネシア国内の ANTARA 社のような新聞社が発行する新聞から 50 万語のテキストを抽出し、それをプロの翻訳者が翻訳し、最後に言語学者や言語処理の専門家によってチェックして作成された。このコーパスはカナダの International Development Research Centre（IDRC）がファンドする PAN Localization Project において SMT の研究のために活用されたもので、インドネシアでの活動は BPPT によって統括された。ドメインとしては、国際、経済、スポーツ、科学の４分野が含まれている。

・BTEC-ATR[232] [233]

ATR が立ち上げた音声翻訳基盤技術の共同研究コンソーシアム A-STAR の中で開発されたコーパス。BTEC とは Basic Travel Expression Corpus の略で、旅行会話で使用される基本的な表現を収集したコーパスである。対訳数は 15.3 万。ATR が保有する英語の単言語コーパスを人手でインドネシア語に翻訳することで作成された。ELRA にて入手可能である。

・ANTARA Corpus[234]

インドネシアの国営新聞社である ANTARA 社が開発したコーパス。同社によるインドネシア語と英語のニュース記事をもとに開発された。テキストは、2000 年から 2007 年にかけて発信された政治、経済、国際・国内ニュース、スポーツ、科学、エンターテインメントなど様々なジャンルの記事から抽出されている。コーパスの規模は、25 万文対、語数で 250 万語である。ELRA にて入手可能である。

・OPUS 対訳コーパス[235]

オープンソースの対訳コーパス構築プロジェクトである OPUS で収集されているコーパスの中にインドネシア語と日本語の対訳コーパスが 2016 年６月現在７つある。それらの合計は文数にして 150 万文対であるが、そのうち 80 万文は映画などのサブタイトルから収集した口語文であり、マニュアルなどから作成した書き言葉のコーパスは残りの 70 万文程度である。また、英語との対訳も 690 万文あるが、そのうち 570 万文はサブタイトルであり、書き言葉は 110 万文ほどである。

232 S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, “Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project.,” in IJCNLP, 2008, pp. 19–24.

233 ELRA-U-W0362:BTEC-ATR Parallel Corpus English – Indonesian, http://universal.elra.info/product_info.php?cPath=42_43&products_id=2204（最終検索日:2016 年 6 月 30 日）

234 Budiono, Hammam Riza, Chairil Hakim: Resource Report: Building Parallel Text Corpora for Multi-Domain Translation System. In proc. of 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009, pp. 92–95, 2009.

235 OPUS – an open source parallel corpus, http://opus.lingfil.uu.se/（最終検索日:2016 年 6 月 30 日）

(2) 単言語コーパス

① 平文コーパス

・Bahara Indonesia Newspapers Collection[236]

情報検索の研究目的に開発された 900 万語の新聞記事コーパスである。ELRA で入手可能である。

② 注釈付きコーパス

・インドネシア大学によるタグ付きコーパス[237]

インドネシア大学が 2014 年に発表したタグ付きコーパス。IDENTIC 対訳コーパスに含まれる対訳のうち、Penn Treebank に含まれていた英文をインドネシア語に翻訳して作成された 2 万 7,325 文対のインドネシア語文にタグをつけたものである。

(3) 対訳辞書

特許以外の主な対訳辞書として以下がある。

・Rekso Translator[238]

Rekso Inovasi 社が 2005 年頃に開発したパソコン用の翻訳支援ソフトであり、これに英語インドネシア語間の対訳辞書が含まれている。辞書の収録語数はインドネシア語英語辞書が 18.9 万語、英語インドネシア語辞書が 23.1 万語である。情報、医学、技術、会計などの専門分野の指定も可能である。2010 年以降更新されていない。

・CICC インドネシア語基本語辞書[239]

（財）国際情報化協力センター（CICC）が 1995 年に作成した辞書の一つに 5 万語のインドネシア語基本辞書が含まれている。辞書のエントリーには対応する英語も付与されているためインドネシア英対訳辞書と見なすことができる。また、コンピュータ、電気関係の専門用語を収録した 2.7 万語の辞書もある[240]。言語資源協会（GSK）から配布されているが、商用利用はできない。

236 ELRA-U-W 0160:Bahasa Indonesia Newspapers Collection, http://universal.elra.info/product_info.php?products_id=1738（最終検索日:2016 年 6 月 30 日）

237 A. Dinakaramani, F. Rashel, A. Luthfi, and R. Manurung, “Designing an Indonesian part of speech tagset and manually tagged Indonesian corpus.,” in IALP, 2014, pp. 66–69.

238 Rekso Translator, https://reksotrans.wordpress.com/category/reksotranslator/（最終検索日:2016 年 6 月 30 日）

239 GSK2006-A-2 CICC インドネシア語基本語辞書, http://www.gsk.or.jp/catalog/gsk2006-a-2/（最終検索日:2016 年 7 月 5 日）

240 GSK2006-A-5 CICC 専門語辞書, http://www.gsk.or.jp/catalog/gsk2006-a-5/（最終検索日:2016 年 7 月 5 日）

・Indodic Dictionary[241]

辞書編集者の Wayne B. Krause（母語は英語）と、インドネシア語が母語の言語学者並びに協力者によるチームによって 10 年近くの歳月をかけて開発された辞書であり、現存する英語-インドネシア語間の辞書としては最も正確で網羅性の高い辞書である。それぞれの辞書は５万語のエントリーを有する。他のインドネシア語-英語辞書と違い、本辞書は接頭辞や接尾辞のほとんどの派生形が収録されており、インドネシア語の学習者にとって理想的な辞書となっている。

(4) インドネシア語辞書

単語辞書として以下がある。

・KBBI[242]

「インドネシア語の偉大な辞書」という意味の *Kamus Besar Bahasa Indonesia*（KBBI）と名付けられた、インドネシア教育省の言語センター（Language Center of the Indonesian Department of Education）によって開発された公式辞書。インドネシア語に取り入れられる借用語の標準形を定める辞書ともなっている。1988 年に発行された初版は 6.2 万語であったが、1991 年の第２版では 7.2 万語、2000 年の第３版では 7.8 万語と増加している。最新の第４版は 2008 年にリリースされたもので９万語が収録されている。オンラインでの利用も可能である。

・Glosarium[243]

上で述べた KBBI と同様、インドネシア教育省の言語センター（Language Center of the Indonesian Department of Education）によって開発された Web アプリケーションで、医学や薬学、獣医学、農業、漁業、林業など種々の応用科学の分野での字訳を支援する目的で開発されたものである。

・Kamus.net[244]

Kamus.net は、STANDS4 Network 社が提供するオンライン辞書サービスである。インドネシア語による語義説明の他、英語の訳語も表示されるため、対訳辞書として使うこともできる。Wikipedia のように、辞書の開発にユーザが参加することも可能である。

241 IndoDic Online, http://indodic.com/index.html（最終検索日:2016 年 6 月 30 日）
242 http://kbbi.kemdikbud.go.id/
243 Glosarium, http://badanbahasa.kemdikbud.go.id/glosarium/（最終検索日:2016 年 6 月 30 日）
244 Kamus.net, http://www.kamus.net/（最終検索日:2016 年 6 月 30 日）

(5) シソーラス

・Indonesian thesaurus Dictionary[245]

Eko Endarmoko によって編集され 2006 年に Gramedia Pustaka Utama Publisher から出版されたシソーラスで、見出し語は 1.6 万語であり、その見出し語の下に同義語が収録されている。ここでの同義語には、意味のニュアンスが違っているもののほか、方言による違い（地域の違いによるいわゆる方言の他、社会的集団に固有な社会方言）なども含まれている。

6.3.3 インドネシア語処理のための要素技術

(1) 形態素解析

主な形態素解析ツールとして以下がある。

・IndMA[246] [247]

IndMA はインドネシア大学の Pisceldo らによって開発された形態素解析器である。下の処理例に示すとおり、インドネシア語の特徴である接辞による派生語に対して、語根と接辞への分割処理を行うことができる。2008 年に公開されたツールが現在も同大学のホームページからダウンロードできる[247]が、利用条件に関する記載はない。

例 18　Membaca → mem + baca （read）

・MorpInd[248] [249]

IDENTIC Corpus を開発したプラハ・カレル大学による形態素解析器である。上で述べた IndMA よりも派生の解析能力が高いのが特徴である。ツールは公開されているが商用利用はできない。

---

245 Tesaurus Bahasa Indonesia, http://www.gramediapustakautama.com/books/80319/detail（最終検索日:2016 年 6 月 30 日）

246 F. Pisceldo, R. Mahendra, R. Manurung, and I. W. Arka, “A two-level morphological analyser for the indonesian language,” in Australasian Language Technology Association Workshop 2008, 2008, vol. 6, pp. 142–150.

247 ツールのダウンロード用データ, http://bahasa.cs.ui.ac.id/tools/MorphologicalAnalyzerIndonesia.zip（最終検索日:2016 年 6 月 30 日）

248 S. D. Larasati, V. Kuboň, and D. Zeman, “Indonesian morphology tool（morphind）: Towards an indonesian corpus,” in Systems and Frameworks for Computational Morphology, Springer, 2011, pp. 119–129.

249 MorphInd: Indonesian Morphological Analyzer, http://septinalarasati.com/work/morphind/（最終検索日:2016 年 6 月 30 日）

・Indonesian Stemmer[250] [251]

Porter Stemmer[252]のアルゴリズムに基づいてアムステルダム大学の Institute for Logic, Language and Computation で開発されたインドネシア語のステマーである。入力された語を語根と接辞に分解する。

・Indonesian Lemmatizer[253] [254]

インドネシアのビナ・ヌサンタラ大学コンピュータ科学研究所で開発された Lemmatizer である。このツールも上で述べたステマー[250]のアルゴリズムを用いているが、処理の目的が見出し語化（lemmatization）となっている。

・POS Tagger（HMM-based）[255] [256]

バンドン工科大学で開発された HMM（隠れマルコフモデル）に基づく品詞タガーである。接辞木（affix tree）と KBBI から取り込んだ lexicon により未知語を含む文に対するタグ付け精度が大きく改善し、30%の未知語を含む場合でも 91.30%の精度が得られたと報告されている[255]。ツールは GitHub にて公開されているが商用利用はできない[256]。

(2) 構文解析

主な構文解析（係り受け解析・依存解析を含む）のツールとして以下がある。

・The Indonesian Mind Map Generator[257] [258]

バンドン工科大学で開発されたインドネシア語の Mind Map Generator の処理の中で行われている自然言語理解の処理モジュールとして、インドネシア語の構文解析器が含まれている。他にも品詞タガーや意味解析器も含まれている。

---

250 F. Z. Tala, “A study of stemming effects on information retrieval in Bahasa Indonesia,” Inst. Log. Lang. Comput. Universeit Van Amst., 2003.

251 Porter Stemmer for Bahasa Indonesia, https://github.com/apraditya/indonesian_stemmer（最終検索日:2016 年 6 月 30 日）

252 M. F. Porter, “An algorithm for suffix stripping,” Program, vol. 14, no. 3, pp. 130–137, 1980.

253 D. Suhartono, D. Christiandy, and R. Rolando, “Lemmatization Technique in Bahasa: Indonesian Language,” J. Softw., vol. 9, no. 5, May 2014.

254 Lemmatizer for indonesian language , https://github.com/davidchristiandy/lemmatizer（最終検索日:2016 年 6 月 30 日）

255 A. F. Wicaksono and A. Purwarianti, “HMM based part-of-speech tagger for Bahasa Indonesia,” in Fourth International MALINDO Workshop, Jakarta, 2010.

256 POS Tag for Bahasa Indonesia , https://github.com/andryluthfi/indonesian-postag（最終検索日:2016 年 6 月 30 日）

257 A. Purwarianti, A. Saelan, I. Afif, F. Ferdian, and A. F. Wicaksono, “Natural Language Understanding Tools with Low Language Resource in Building Automatic Indonesian Mind Map Generator,” Int. J. Electr. Eng. Inform., vol. 5, no. 3, p. 256, 2013.

258 http://mindmap.kataku.org リンク切れ

・Indonesian Language Symbolic Parser[259] [260]

インドネシア大学で開発されたパーザで、2002 年に発表された。文脈自由文法による規則を使って解析を行うルールベースシステムであり、国際 SIL（SIL International）が提供する構文解析器にて動作する。

・Indonesian Language Semantic Analyzer[261]

インドネシア大学で開発された意味解析器で、2007 年に発表された。意味表現形式として一階述語論理を用いており、処理系として PROLOG を用いている。

(3) 述語項解析

・MorpInd[248]

6.3.3(1)で述べた MorpInd は述語項解析が可能である。

・Annotated Disjunct（ADJ）[262]

マレーシアのペトロナス技術大学によって開発された解析器。Link Grammar に基づく英語の構文解析と英語-インドネシア語注釈辞書を組み合わせて実現している。

(4) 統計的機械翻訳システム

・英語-インドネシア語の SMT システム [263]

統計的機械翻訳システムのツールとして著名な Moses[264] [265]のデコーダと SRI による言語モデルツールの SRILM、アライメントツールとして広く使われている GIZA++を用いて、BPPT（Agency for the Assessment and Application of Technology:インドネシア技術応用評価庁）が中心となって開発したもの。開発は対訳コーパスの項で記載した PAN Localization Project の中で行われた。文献[263]によれば、商用のシステムと同等の性能、一部では Google 翻訳よりも良い結果が得られたとの記載があるが、具体的な精度は示されていない。

---

259 Joice, ‘Pengembangan lanjut pengurai struktur kalimat bahasa indonesia yang menggunakan constraint-based formalism,’ Undergraduate Thesis, Faculty of Computer Science, University of Indonesia, 2002, call num- ber: SK-0487.

260 ツールのダウンロード用データ, http://bahasa.cs.ui.ac.id/tools/SymbolicParser.zip（最終検索日:2016 年 6 月 30 日）

261 ツールのダウンロード用データ, http://bahasa.cs.ui.ac.id/tools/SemanticAnalyzer.zip（最終検索日:2016 年 6 月 30 日）

262 T. B. Adji, B. Baharudin, and N. B. Zamin, “Annotated Disjunct in Link Grammar for Machine Translation,” in Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on, 2007, pp. 205–208.

263 M. Adriani and H. Riza, “Research Report on Local Language Computing: Development of Indonesian Language Resources and Translation System,” Ref No PANL10nAdmnRR001 PAN Localization Proj., 2008.

264 P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, and R. Zens, “Moses: Open source toolkit for statistical machine translation,” in Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, 2007, pp. 177–180.

265 Moses, http://www.statmt.org/moses/（最終検索日:2016 年 6 月 30 日）

6.4　中国語

6.4.1　中国語の特徴

(1) 文字

中国語の表記に使われている文字には二種類ある。一つは中国本土で用いられている簡体字で、もう一つは台湾や香港で用いられている繁体字である。どちらにも日本語で用いられている漢字と共通の字もあるが、異なっているものも少なくない。そのため日ごろから漢字を使っている日本人であっても、中国語の学習者でなければ簡体字、繁体字の漢字を見て日本語の対応する漢字を想起するのは難しい。

表６．４．１　日本語の漢字との比較

日本語の漢字	簡体字	繁体字
亜	亚	亞
図	图	圖
広	广	廣

(2) 分かち書き

下の例１に示すように、中国語の文では英語のように空白を用いた分かち書きは行われない。しかも、漢字だけでなくひらがなやカタカナを用いる日本語と違い、中国語で用いられる文字は原則漢字だけである。読みやすくするために読点が用いられたり、特許文書のような技術文献では数値やアルファベットが使われたりすることもあるが、基本的には下の例のような漢字が連続した文字列を単語に分割することになる。我々人間にとっても難しいのと同様、中国語文をコンピュータで処理するのは、日本語での場合に比べ相対的に難しいと言える。

例１　槽内有小镊子。（溝内にピンセットがある。）

↓

槽/内/有/小镊子/。

(3) 語順

中国語はいわゆる SVO 言語である。そのため、動詞が目的語を取る場合、例２のように目的語は動詞の後ろに置かれる。また、英語の前置詞に相当する介詞という品詞の語も、英語のように使われる。そういう点から、中国語は英語に近い言語であると評されることもあるが、その一方で英語の関係節のような修飾の場合は、日本語と同様、修飾句は被修飾語の前に置かれ、日本語に近い性質も持っている。

例２　本发明提供下式的化合物。

↓

本发明（本発明）　提供（提供する）　下式（下式）　的（の）　化合物（化合物）

(4) 中国語処理を難しくしている中国語の特徴

上で述べた分かち書きの項で、分かち書きされておらず漢字が連続しているという中国語の特徴が中国語処理が難しい要因であると述べたが、日本語と比べた時に他にも以下のような特徴がある。

①　語尾がないこと

日本語では、動詞にしろ、形容詞・形容動詞にしろ、活用する語尾を持っている。それに対し中国語の用言にはそのような活用する語尾が存在しない。しかも、そのような語が名詞として用いられたり用言として用いられたりする。

例えば、下の例３には「安装」という語が３回現れている。文頭に近い方の語と文末の語（赤色の語）は、直後の名詞「线」、「部」とともに複合名詞を構成する。一方、２つ目の語（青色の語）は動詞である。このように２つの「安装」は違う役割を果たすが、動詞の「安装」には日本語のような語尾がついてないので他の２つと見た目は全く同じであり、二種類の「安装」が文中で果たす役割の違いを単語の形態から判定することは難しい。この特徴は、中国語の構文解析そして中国語を原言語とする機械翻訳の精度向上を難しくしている原因の一つである。

例３　利用安装线５将链齿１安装在链齿安装部Ａ。

（取付糸５によってエレメント１をエレメント取付部Ａに取り付ける。）

②　格助詞が少ないこと

例４に示すように、中国語文では、日本語訳文にある「に」・「が」のような格助詞が存在しない。日本語文を解析する際は、「に」と「が」のような格助詞が重要な手がかりになり、構文要素の切れ目の印になるだけでなく、構文要素が果たす文法的な役割や意味的な役割を表す手がかりになっている。中国語文では、このような格助詞が少ないため、中国語文の解析においては、構文要素の切れ目やその文法的な役割の判定は、日本語より難しくなる。

例４　外壳内安装有电路板,（ケース内に回路基板が取り付けられ,）

③　外来語表記の問題

日本語では外来語を音訳してカタカナ表記する場合が少なくない。翻訳としては安易

な方法であり現代日本語でカタカナ語が氾濫する要因ともなっているが、カタカナ表記の部分は通常漢字やひらがなの部分とは意味的に分離できる場合が多いため、言語処理用の辞書に登録されていない未知語が含まれている場合でも、カタカナ部分を切り出して名詞として処理すれば構文解析は問題なく行えることが多く、言語処理の観点からは好ましい表記方法となっている。

それに対し中国語では外来語を意訳するケースが多い。例えば、下に例を示した「百草枯」は英語「Paraquat」の訳である。この Paraquat は除草剤の一種であり、中国語での訳は「百草（いろいろの種類の草)」と「枯（枯れる)」という２つの単語を組み合わせて「どんな草でも枯れさせる」という意味を表している。このような意訳された語は単語分割されると、もともとの意味が失われるだけでなく、原文の構文構造を本来の構造とは全く異なるものへと導く原因となる。すなわちこの例であれば、「百草枯」の部分を名詞として解釈すべきところが、「いろいろの種類の草が枯れる」という節として解釈されることになり、結果として正しく翻訳は望めなくなる。このように、意訳された外来語を正しく切り出すことは難しく、外来語由来の新規の概念や商標などが用いられる可能性が高い中国語特許文書の機械翻訳においては重要な課題の一つである。

例５　百草枯（パラコート）

④　単語の定義が難しいこと

中国語の漢字は意味を表すため、単語の定義が難しい。これは、単語分割の基準を統一しにくい原因となる。例えば、ペンシルベニア州立大学の研究チームが開発した Penn Chinese Treebank（CTB）と北京大学が構築した Peking University Treebank（PKU）の単語分割基準はそれぞれ違う[266] [267]。例えば、例６にある「擦干」は、CTB の基準では１つの単語であるが、PKU の基準では「擦」と「干」に分ける必要がある。このような単語単位の揺れは、各種タグ付きコーパスを開発する際に問題となる。

例６　然后，用毛巾充分擦干水分。（そして、タオルで水分をよく拭き取った。）

6.4.2　中国語処理のための言語リソース

(1) 対訳コーパス

中国語の特許対訳コーパスとして現在公開されているものは、特許庁と NICT が共同で開発した対訳コーパスがある。

---

266 Fei Xia（2000）. The Segmentation guidelines for Chinese Treebank Project. Technical Report. IRCS00-06, University of Pennsylvania.

267 俞士汶,段慧明,朱学锋,孙斌（2002）. 北京大学現代中国語コーパスにおける構築基準（北京大学现代汉语语料库基本加工规范）. Journal of Chinese Information Processing, 16（5）: 51-66.

・JPO 中日対訳コーパス

「JPO 中日対訳コーパス」は、中国語と日本語の対応する公開特許公報の対（パテントファミリー）若しくは中国語の公開特許公報の要約部分とそれを特許庁が日本語に翻訳した和文抄録データの対をもとに、日本国特許庁（JPO）と NICT が作成したデータである。データ範囲 2005 年から 2013 年（約１億 3,285 万文対）が研究目的で高度言語情報融合フォーラム（ALAGIN）[268]から、2005 年～2014 年（約１億 7,318 万文対）が特許庁[269]から公開されている。利用条件は共に研究目的である。

特許以外の対訳コーパスには以下のようなものがある。

・Chinese-English Sentence-Aligned Bilingual Corpus（英汉双语平行语料库）[270]

Datum Data Co.,Ltd.（点通数据有限公司）が 2006 年 12 月から 2009 年にかけて構築した 2,000 万文対の対訳コーパスである。Chinese Linguistic Data Consortium で配布されている。最新の更新時間は 2009 年 12 月３日で、研究目的のライセンス（中国国内：0.003RMB/文；国外：0.021RMB/文）のほか、商用目的のライセンス（中国国内：0.015RMB/文；国外：0.105RMB/文）もある。

・Chinese-English Sentence-Aligned Bilingual Corpus（中英句子级对齐双语语料库）[271]

中国科学院計算研究所と自動化研究所が 2003 年開発した 21 万文対の中英文対訳コーパスである。Chinese Linguistic Data Consortium で配布されている。最新の更新時間は 2010 年 11 月 11 日で、研究目的で使用可能である（有料。中国国内：3,000RMB；国外：2.1 万 RMB)。

・Chinese-English/Chinese-Japanese parallel corpora（汉英/汉日双语语料库）[272]

北京大学計算言語学研究所が 2003 年開発した中英・中日対訳コーパスである。このコーパスには中英文対訳コーパス 20 万文対、中英単語対訳コーパス１万文対、中日文対訳コーパス２万文対が含まれている。Chinese Linguistic Data Consortium で配布されている。最新の更新時間は 2010 年２月 24 日で、研究目的で使用可能である（有料。中国国内：2,000RMB（CJ）；6,000RMB（CE）；国外：１万 RMB（CJ）；２万 RMB（CE））。

---

[268] JPO コーパス概要, https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html（最終検索日:2016 年 7 月 1 日）

[269] https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyo_dictionary.htm（最終検索日:2016 年 7 月 20 日）

[270] 英汉双语平行语料库, http://www.chineseldc.org/resource_info.php?rid=141（最終検索日:2016 年 7 月 1 日）

[271] 中英句子级对齐双语语料库, http://www.chineseldc.org/resource_info.php?rid=50（最終検索日:2016 年 7 月 1 日）

[272] 汉英/汉日双语语料库, http://www.chineseldc.org/resource_info.php?rid=40（最終検索日:2016 年 7 月 1 日）

(2) 対訳辞書

・JPO 中日対訳辞書

「JPO 中日対訳辞書」は、中国語と日本語の対応する公開特許公報の対（パテントファミリー）若しくは中国語の公開特許公報の要約部分とそれを特許庁が日本語に翻訳した和文抄録データの対をもとに、日本国特許庁（JPO）が作成したデータである。(6.7.1(4)の日本語の対訳辞書を参照)

・EDR 日英中翻訳辞書[273]

日本電子化辞書研究所（EDR）が 1986 年から 1994 年に作成した辞書で、英日対訳辞書が基本語 16 万語、日英対訳辞書が基本語 23 万語、日中対訳辞書が基本語 23 万語を収録している。商用利用可能である。

(3) 品詞付けコーパス

・Chinese POS Tagged Corpus（分词词性标注语料库）[274]

中国山西大学が 2002-2003 年開発した 500 万漢字規模の品詞付けコーパスである。Chinese Linguistic Data Consortium で配布されている。最新の更新時間は 2008 年６月 11 日で、研究目的でも（中国国内：8,000RMB；国外：4.2 万 RMB)、商用目的でも使用できる（中国国内：５万 RMB；国外：8.4 万 RMB)。

・北京大学現代中国語コーパス [267]

北京大学計算言語学研究所が開発した 2,700 万漢字の中国語コーパスである。単語分割、品詞情報のほか、人名、地名などの専門用語の情報も含まれている。

(4) 中国語ツリーバンク

・Chinese Penn Treebank（CTB）[275]

最初はペンシルベニア州立大学の研究チームが開発した句構造文法情報付き言語リソースであった。最新のバージョンは Brandeis University によって更新された Chinese Treebank 8.0 で、米 Linguistic Data Consortium（LDC）から配布されている[276]。更新時間は 2013 年 11 月 15 日である。使用料金は LDC 会員の場合は無料、非会員の場合は＄300.00 である。

[273] 日中対訳辞書およびＥＤＲ電子化辞書 Ver4.0 について, https://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html（最終検索日:2016 年 7 月 1 日）

[274] 分词词性标注语料库, http://www.chineseldc.org/resource_info.php?rid=31（最終検索日:2016 年 7 月 1 日）

[275] Xue, N. and Xia, F.（2000）. “The Bracketing Guidelines for the Penn Chinese Treebank.” Technical Report. University of Pennsylvania.

[276] Chinese Treebank 8.0, https://catalog.ldc.upenn.edu/LDC2013T21（最終検索日:2016 年 7 月 1 日）

・Peking University Treebank（PKU）[277]

中国北京大学が構築した句構造文法情報付き言語リソースである。

・Tsinghua Chinese treebank（TCT）（汉语句法树库）[278]

中国語清華大学の計算機系知能技術及びシステム国家重点実験室が 1998 年から 2003 年にかけて、開発した句構造文法情報付き言語リソースである。規模は 100 万漢字。最新更新時間は 2014 年 12 月 30 日で、研究目的でも（１万 RMB)、商用目的でも使用できる（10 万 RMB）[279]。

・Chinese Dependency Treebank 1.0[280]

中国ハルピン工業大学が開発した依存文法情報付き言語リソースである。最新更新時間は 2012 年５月 16 日で、サイズは 49,996 文（902,191 語)。米 Linguistic Data Consortium（LDC）から配布されている。使用料金は LDC 会員の場合は無料、非会員の場合は＄300.00 である。

・Sinica Treebank[281] [282]

台湾中央研究院が Information-based Case Grammar の考え方に基づき、開発した言語リソースである。最新バージョン Sinica Treebank Version 3.0 は 61,087 文（361,834 単語）規模。研究目的での使用のみ可能。

(5) 中国語述語項付きツリーバンク

・Chinese Proposition Bank 3.0[283]

Chinese Proposition Bank 3.0 は、Chinese Treebank 7.0[284]における 187,731 単語に述語項構造情報を加えたもので、米 Linguistic Data Consortium（LDC）から配布されている。最新更新時間は 2013 年７月 15 日である。使用料金は Linguistic Data Consortium 会員の場合は無料、非会員の場合は＄300.00 である。

277 Yu, S. et al.（2002）. “The Basic Processing of Contemporary Chinese Corpus at Peking University Specification.” Journal of Chinese Information Processing, 16（5）.

278 Zhou, Q.（2004）. “Annotation Scheme for Chinese Treebank.” Journal of Chinese Information Processing, 18（4）.

279 汉语句法树库, http://www.chineseldc.org/resource_info.php?rid=33（最終検索日:2016 年 7 月 1 日）

280 Chinese Dependency Treebank 1.0, https://catalog.ldc.upenn.edu/LDC2012T05 （最終検索日:2016 年 7 月 1 日）

281 Chen Keh-Jiann, Chu-Ren Huang, Li-Ping Chang, Hui-Li Hsu.（1996）. “Sinica Corpus: Design Methodology for Balanced Corpra.” Proceedings of the 11th Pacific Asia Conference on Language, Information, and Computation（PACLIC II）, SeoulKorea, pp.167-176.

282 中文句結構樹資料庫簡介, http://turing.iis.sinica.edu.tw/treesearch/（最終検索日:2016 年 7 月 1 日）

283 Chinese Proposition Bank 3.0, https://catalog.ldc.upenn.edu/LDC2013T13（最終検索日:2016 年 7 月 1 日）

284 Chinese Treebank 7.0, https://catalog.ldc.upenn.edu/LDC2013T107（最終検索日:2016 年 8 月 2 日）

(6) 中国語文法規則リソース（hand-crafted）

・Chinese Sentence Structure Grammar（CSSG）[285]

中国人研究者が中国語の孤立語としての文法特徴に着目し、その深層文法制約を捉えようとする文法枠組み Sentence Structure Grammar（SSG）を提案し、それに基づき、人手で開発した文法規則リソースである。文法規則の網羅率が 94.2%に達したことが検証された。

(7) 中国語文法辞書

・中国語文法情報辞書（现代汉语语法信息词典）[286]

中国北京大学計算言語学研究所が 1990 年代に開発した文法情報付きの辞書である。

・中国語文法情報辞書（高頻度詞）（现代汉语语法信息词典（高频词））[287]

中国北京大学計算言語学研究所が 2003 年に 2,600 万漢字の新聞コーパスから頻度の高い 2.8 万語を選び、作成した文法情報付きの辞書である。Chinese Linguistic Data Consortium で配布されている。最新更新時間は 2009 年 12 月 3 日で、研究目的でも（中国国内：4,000RMB；国外：2.8 万 RMB)、商用目的でも使用できる（中国国内：1.6 万 RMB；国外：11.2 万 RMB)。

(8) 意味関係辞書

・HowNet[288]

個人の研究者が独自のアイデアに基づき、開発した意味関係辞書である。６万漢字規模で、研究目的で使用可能である。

・Chinese FrameNet[289]

中国山西大学が Fillmore の Frame semantics に基づき、米カリフォルニア大学バークレー校が開発した英語版の FrameNet を参考して構築した 200 文献（フレーム 323、単語 3,947、例文２万）規模の意味関係辞書である。

285 Wang Xiangli, Yi Zhang, Yusuke Miyao, Takuya Matsuzaki, Junichi Tsujii（2013）. Deep Context-free Grammar for Chinese with Broad Coverage. In the Proceedings of SIGHAN-7.

286 Yu, S.（1998）. 现代汉语语法信息词典详解. 清華大学出版社.

287 现代汉语语法信息词典（高频词）, http://www.chineseldc.org/resource_info.php?rid=30（最終検索日:2016 年 7 月 1 日）

288 HowNet HP, http://www.keenage.com/html/e_index.html（最終検索日:2016 年 7 月 1 日）

289 Chinese FrameNet, http://sccfn.sxu.edu.cn/portal-en/home.aspx（最終検索日:2016 年 7 月 1 日）

(9) 単語品詞辞書

・CSSG 単語品詞辞書[290]

中国人研究者が Sentence Structure Grammar（SSG）という文法枠組みに基づき、品詞体系を設計した。さらにこの品詞体系に基づき、30.6 万単語規模の単語品詞辞書を開発した。

6.4.3 中国語における要素技術

(1) 単語分割技術

・ICTCLAS[291]

中国科学院計算技術研究所が開発した中国語形態素解析ツール。分割精度が 97.58%で、スピードは 31.5KB/s である。商用利用可能である。

・Stanford Chinese Word Segmenter[292]

米スタンフォード大学が開発した中国語単語分割ツールである。このツールはオープンソースソフトウェアであり、そのライセンスは GPL に従い、商用可能である。

・CSP（Chinese Word Segmenter and POS Tagger）[293]

NICT が開発した中国語形態素解析ツールで、単語分割の正解率は 95.66%である。文献[293]によれば、このツールは ALAGIN 言語資源サイトを通じてオープンソースソフトウェアとして一般公開される予定との記載があるが、本調査の時点では公開されていない。

・NiuParser[294]

中国東北大学が開発した中国語文解析ツールであり、単語分割もできる。研究目的で無料利用が可能で、商用目的で利用する場合は、ライセンスを受ける必要がある。このツールの正解率は 97.3%、スピードは 45K 字/秒、モデルサイズは 57MB、使用されるメモリサイズは 68MB である[295]。

---

290 王向莉（2015）.NLP に重要なのは、学習データの量なのか、言語学知識体系の質なのか？ -中国語単語分割タスクで検証する-．言語処理学会第 21 回年次大会発表論文集,pp.1068-1071.

291 Introduction to ICTCLAS, http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html（最終検索日:2016 年 7 月 1 日）

292 Stanford Word Segmenter, http://nlp.stanford.edu/software/segmenter.shtml（最終検索日:2016 年 7 月 1 日）

293 風間淳一,王軼謳,川田拓也（2012）.基盤的言語処理ツール.情報通信研究機構季報, Vol. 58 Nos. 3/4.

294 Chinese Syntactic and Semantic Parser NiuParser 1.3.0, http://www.niuparser.com/index.en.html （最終検索日:2016 年 7 月 1 日）

295 Jingbo Zhu , Muhua Zhu , Qiang Wang , Tong Xiao. 2015. NiuParser: A Chinese Syntactic and Semantic Parsing Toolkit. In Proc. of ACL, demonstration session.

(2) 品詞付け技術

・ICTCLAS[291]

中国科学院計算技術研究所が開発した中国語形態素解析ツールである。商用利用可能である。

・Stanford Chinese Word tagger[296]

スタンフォード大学が開発した中国語品詞付けツールで、その正解率は 93.65%である[297]。このツールはオープンソースソフトウェアであり、そのライセンスは GPL に従い、商用可能である。

・CSP（Chinese Word Segmenter and POS Tagger）[293] [298]

NICT が開発した中国語形態素解析ツールで、品詞付けの正解率は 90.51%である。このツールは ALAGIN 言語資源サイトを通じてオープンソースソフトウェアとして一般公開している。

・NiuParser[294]

中国東北大学が開発した中国語文解析ツールであり、品詞付け機能もある。研究目的で無料利用が可能で、商用目的で利用する場合は、ライセンスを受ける必要がある。このツールの正解率は 93.5%、スピードは 38.8K 字/秒、モデルサイズは 185MB、使用されるメモリサイズは 93MB である[295]。

(3) 構文解析技術

・Stanford Chinese parser[299]

スタンフォード大学が開発した中国語構文解析器であり、正解率は 78.8%である。このツールはオープンソースソフトウェアであり、そのライセンスは GPL に従い、商用可能である。

・CNP（A ChiNese dependency Parser）[300]

NICT が開発した中国語係り受け解析器で、正解率は 91.93 である。このツールは ALAGIN 言語資源サイトを通じてオープンソースソフトウェアとして一般公開している。

---

296 Huihsin Tseng, Daniel Jurafsky, Christopher Manning（2005）. Morphological features help POS tagging of unknown words across language varieties. The Fourth SIGHAN Workshop on Chinese Language Processing, 2005

297 Chinese Natural Language Processing and Speech Processing , http://nlp.stanford.edu/projects/chinese-nlp.shtml（最終検索日:2016 年 7 月 1 日）

298 https://alaginrc.nict.go.jp/cnp/（最終検索日:2016 年 8 月 4 日）

299 Roger Levy and Christopher Manning（2003） Is it harder to parse Chinese, or the Chinese Treebank? Proceedings of ACL 2003.

300 CNP - A ChiNese dependency Parser, https://alaginrc.nict.go.jp/cnp/index.html（最終検索日:2016 年 7 月 1 日）

・NiuParser[294]

中国東北大学が開発した中国語文解析ツールであり、構文解析ができる。研究目的で無料利用が可能で、商用目的で利用する場合は、授権される必要がある。このツールの正解率は 83.2%、スピードは 583.3 字/秒、モデルサイズは 243MB、使用されるメモリサイズは 0.98GB である[295]。

(4) 意味解析技術

・NiuParser[294]

中国東北大学が開発した中国語文解析ツールであり、述語項付け機能がある。研究目的で無料利用が可能で、商用目的で利用する場合は、授権される必要がある。このツールの正解率は 68.4%、スピードは 494 字/秒、モデルサイズは 30MB、使用されるメモリサイズは 1.2MB である[295]。

6.5 韓国語

6.5.1 韓国語の特徴

韓国語は、世界の言語の中で日本語と最も類似した言語である。助詞の体系も日本語とよく似ており、語順もほぼ同じである。そのため、特許文書を含め技術文献を翻訳する場合、韓国語の単語を機械的に日本語に置き換えるだけで、ある程度意味の通る日本語になる。表音文字であるハングルを用いて表記されることから日本人は簡単には読めないが、漢語由来の言葉や外来語由来の言葉がハングルではなく元の漢字やアルファベットのまま表記されていれば、日本人ならば、ほぼそのまま理解できる言語であると言える。

(1) 韓国語の文字

韓国語の文献には、「ハングル」と「アルファベット」が使われる。同音異義語を補足するため、まれにカッコ書きで「漢字（旧字体・繁体字)」が記されることがある。ここでは、韓国語固有のハングルについて簡単に記す。

ハングルには 19 種類の子音記号と 21 種類の母音記号があり、それらを組み合わせた 399 文字（=19×21）が基本形となる。

（子音記号）

ㄱ（g） ㄴ（n） ㄷ（d） ㄹ（l/r） ㅁ（m） ㅂ（b） ㅅ（s） ㅇ（無音） ㅈ（j）
ㅊ（ch） ㅋ（k） ㅌ（t） ㅍ（p） ㅎ（h） ㅉ（jj） ㄲ（kk） ㄸ（tt） ㅃ（pp） ㅆ（ss）

（母音記号）

ㅏ（a） ㅑ（ya） ㅓ（eo） ㅕ（yeo） ㅗ（o） ㅛ（yo） ㅜ（u） ㅠ（yu） ㅡ（eu） ㅣ（i）

ㅐ（ae） ㅒ（aye） ㅔ（e） ㅖ（ye） ㅘ（wa） ㅙ（wae） ㅚ（oe） ㅝ（wo） ㅞ（we） ㅟ（wi） ㅢ（eui）

（単語例）

기구（gi-gu ： 機構）、기기（gi-gi ： 機器）、시야（si-ya ： 視野）、
표시（pyo-si ： 表示）、도체（do-che ： 動体）、회로（hoe-lo ： 回路）、
미디어（mi-di-eo ： メディア）

さらに、前項「子音」＋「母音」を組合せた文字の下に「子音記号」がつく文字パターンもある。この子音記号のことを「パッチム」といい、後述の活用形に影響を与える。パッチムには、二つの子音記号を組合せたものを含め 27 パターンあるため、理論上は 11,172（=399×（1+27））文字存在するが、その中で実際に使われる組合せは 2,305 文字と言われている。

（パッチムが含まれる単語例）

로봇（lo-bos ： ロボット）、웹（web ： ウェブ）

(2) 分かち書き（語節）

韓国語は英語と同じく空白を用いて分かち書きされる。分かち書きされる箇所は日本語における文節の切れ目と考えればよい。意味が通る最小単位で区切られるため、助詞は単独で表記されない。この分かち書きの単位を語節という。

例１：[301]

OLED는 형광 또는 인광 유기물 박막에 전류를 흘리면
OLED は/蛍光/または/リン光/有機物/薄膜に/電流を/流せば

전자와 정공이 유기물 층에서 결합하면서
電子と/正孔が/有機物/層で/結合しつつ

빛이 발생하는 원리를 이용한 자체 발광형 디스플레이를 말한다.
光が/発生する/原理を/利用した/自体/発光型/ディスプレイを/言う。

(3) 活用

時制や複文の接続などは、日本語と同じく動詞・形容詞を活用することで行う。日本語「する」の活用（例：して、するが、したが、するものの、するであろうが etc…）と同

[301] 出典：http://terms.naver.com/entry.nhn?docId=1260888&cid=40942&categoryId=32382,（最終検索日:2016 年 7 月 1 日）

じ現象が韓国語にも存在する。日本語の「する」に相当する語尾「하다」には、過去連体形「した」を意味する「한」がある[302]。

例２：　발광하다　（発光する）
　　발광한　디스플레이　（発光したディスプレイ）

(4) 語順

例１の例文からも分かるとおり、韓国語の語順は日本語とほぼ同じで SOV 言語のグループに属し、修飾句は被修飾句の前に来るといった大まかな特徴のみならず、助詞の体系もほぼ同じであることから、技術文献であれば、韓国語の単語を機械的に日本語に置き換えるだけで、ある程度意味の通る日本語になる。

したがって、韓国語と日本語間の機械翻訳は、英語や中国語を日本語に翻訳する場合と比べ、現状の RBMT 方式の機械翻訳であっても実用に耐えうるシステムを比較的容易に構築できる。なお、韓日翻訳時に訳語選択が必要な同音異義語については、できるだけ複合語として辞書登録するなどして訳語選択精度を改善する。同音異義語の例としては「전기（jeon-gi:電気/電機/電器）」等がある。

6.5.2　韓国語処理のための言語リソース

(1) 韓国語と外国語間の対訳コーパス

特許をもとにした対訳コーパスとして以下がある。

・JPO・NICT 韓日対訳コーパス（約 8346 万文対）

日本国特許庁と NICT の共同研究でパテントファミリーの公開特許公報から作成された対訳コーパスで、高度言語情報融合フォーラム（ALAGIN）から研究目的で配布[303]されている。

また、特許以外の主な対訳コーパスとして以下がある。

・Sejong コーパス[304]（2000-2007 年）

国立国語院が韓国語研究基盤を確保するため 21 世紀世宗計画（韓国語コーパス構築プロジェクト）で作成した、英韓・日韓形態素解析情報付きの対訳コーパスである。

---

302 活用や助詞の付与状況を解析するプログラムが Web で公開されている。須賀井義教,「MeCab による韓国語の形態素解」,http://porocise.sakura.ne.jp/korean/mecab/analyzer.html,（最終検索日:2016 年 7 月 1 日）

303 https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html

304 https://ithub.korean.go.kr/user/main.do（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

Creative Commons ライセンス[305]で商用目的では利用できない。

対訳コーパスのデータ量は英韓がソース 600 万 6,253 語、形態素解析 130 万 9,330 語[306]、日韓がソース 115 万 8,248 語、形態素解析 29 万 7,953 語である。

・日韓対訳コーパス検索システム NARA サービス[307]

21 世紀世宗計画の韓日対訳コーパスを検索するサービス。無料であるが、研究及び教育目的にのみ利用可能である。

・KAIST（Korea Advanced Institute of Science and Technology）言語リソース銀行コーパス[308]

KAIST の SWRC 研究センターが、Semantic 関連の技術研究に必要な言語リソースを普及するために作成した対訳コーパスである。研究目的であれば無料で利用できるが、商用目的では別途契約（有料）が必要である。

① 中英韓対訳コーパス（2000 年）：６万文セット
② 中英対訳コーパス（2005 年）：６万文セット
③ 中韓対訳コーパス（2005 年）：６万文セット
④ 英韓対訳コーパス（2005 年）：６万文セット
⑤ 韓英日新聞対訳コーパス（2005 年）：1,791 ファイル

・機械翻訳用国文[309]（2008 年）

KIPRIS（韓国特許技術情報センター）が、海外向けに、韓国の特許情報を英語文に加工した KPA を、再加工して機械翻訳した抄録データある。

データ提供期間は 2008 年～現在までで、2014 年 1 月時点で 75 万文献である。無料であるが、商用目的の場合は事前協議及び許諾が必要である。

---

305 https://ja.wikipedia.org/wiki/クリエイティブ・コモンズ・ライセンス
Creative Commons license は著作権のある著作物の配布を許可するパブリック・ライセンスの一つである。

306 語節は韓国語の分かち書きの単位。単語に日本語の助詞を続けたような単位。
http://krdic.naver.com/detail.nhn?docid=26146800

307 ソン サンホン, NARA System（ver. 2.0）, http://corpus.mireene.com/nara.php（最終検索日:2016 年 6 月 10 日）

308 Semantic Web Research Center, Bank of Resource for Language and Annotation, http://semanticweb.kaist.ac.kr/org/bora/index.php（最終検索日:2016 年 6 月 10 日）（Semantic Web Research Center は言語リソースを管理・普及する KAIST 傘下国家機関）

309 KIPRIS, BULK DATA 기계번역용 국문초록 （機械翻訳のハングル抄録）, http://plus.kipris.or.kr/portal/data/service/DBII_000000000000025/view.do?pageIndex=4&menuNo=200101&kppBCode=&kppMCode=&kppSCode=&subTab=SC003&entYn=&clasKeyword=（最終検索日:2016 年 6 月 10 日）

・韓国特許の英語抄録データ[310]（1979 年）

KIPRIS（韓国特許技術情報センター）が、海外向けに、韓国特許を英語文で加工して作成した明細書と図面データである。

データ提供期間は 1979 年～現在までで、2014 年 1 月時点で 199 万文献である。月単位でデータを更新し、有料で研究、商用目的で利用できる。データの利用料は本年度：282.6 万 KRW、過去分：1,554.3 万 KRW である。

・韓国大手新聞社の対訳データ

韓国の三大新聞社の記事を、人手で英語、日本語、中国語に翻訳したものをもとに作成した対訳コーパスであり、各言語あたり 100 万文対程度のコーパスを含む。記事は政治、経済、社会、スポーツ、娯楽等多岐にわたり、調査・研究目的のほか、商用目的での利用も可能。日本国内では株式会社高電社を通じて有料で販売されている。

・名詞／名詞連語の韓英対訳パターン DB（2004 年）

ETRI（韓国電子通信研究院）が 2004 年から知識情報資源管理事業で構築した DB である[311]。

「名詞／名詞連語の韓英対訳パターン」は一般文書（新聞、Web 文書）等から、使用頻度が高い連語情報を名詞／名詞及び名詞／動詞関係に区分して抽出した「名詞／名詞連語の韓英対訳パターン」をデータとして含む。名詞／名詞連語の韓英対訳パターンは、対訳語の変化が起こす多義語、又は個別単語対訳語との単純な組み合わせとは違う翻訳になる複合名詞を対象にしたデータである。

名詞／動詞連語韓英対訳パターン 93 万対、名詞／名詞連語韓英対訳パターン 13 万対からなる。

有料で研究、商用目的で利用できる。

・韓英連語対訳パターン DB（2005 年）

ニュース、新聞記事をもとにしたコーパスから、「名詞+副詞格助詞+用言」のパターンを抽出したデータである。

ユニークな「名詞+副詞格助詞+用言」が 50 万組ある。有料で研究、商用目的で利用できる。

・韓英中日口語体対訳コーパス（2011 年）

旅行会話をもとにして作成した口語体コーパスデータである。

---

310 KIPRIS, BULK DATA 한국특허영문초록 （KPA）（韓国特許英文抄録）, http://plus.kipris.or.kr/portal/data/service/DBII_000000000000024/view.do?pageIndex=3&menuNo=200101&kppBCode=&kppMCode=&kppSCode=&subTab=SC003&entYn=&clasKeyword=（最終検索日:2016 年 6 月 10 日）

311 https://etri.re.kr/tcenter/db/db_form.jsp（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

対訳コーパスの組み合わせが韓英：10 万文対、韓中：10 万文対、韓日：10 万文対である。有料で研究、商用目的で利用できる。

・韓スペイン口語体対訳コーパス（2013 年）

生活・旅行・ビジネス会話をもとにして作成した口語体コーパスデータである。

韓国語とスペイン語の対訳コーパスが 20 万文対である。有料で研究、商用目的で利用できる。

・韓仏口語体対訳コーパス（2014 年）

生活・旅行・ビジネス会話をもとにして作成した口語体コーパスデータである。

韓国語とフランス語の対訳コーパスが 10 万文対である。有料で研究、商用目的で利用できる。

・韓独ロシアベトナム口語体対訳コーパス（2015 年）

生活・旅行・ビジネス会話をもとにして作成した口語体コーパスデータである。

対訳コーパスの組み合わせが韓独：５万文対、韓ロシア：５万文対、韓ベトナム：５万文対である。有料で研究、商用目的で利用できる。

(2) 対訳辞書

・日韓特許技術用語翻訳辞書[312]

KIPRIS が、機械翻訳の精度向上のために作成したデータである。

語数は 2015 年 11 月 25 日更新で 48 万語である。無料であるが、商用目的の場合は事前協議及び許諾が必要である。

・英韓特許技術用語翻訳辞書[313]

KIPRIS が、機械翻訳の精度向上のために作成したデータである。

語数は 2015 年 11 月 25 日更新で、84 万語である。無料であるが、商用目的の場合は事前協議及び許諾が必要である。

312 KIPRIS, 일한 특허기술용어 번역사전 （日韓特許技術用語翻訳辞書）, http://plus.kipris.or.kr/portal/data/main/DBII_000000000000023/view.do?menuNo=210103&kppBCode=&kppMCode=&kppSCode=&subTab=&entYn=&clasKeyword=（最終検索日:2016 年 6 月 10 日）

313 KIPRIS, 영한 특허기술용어 번역사전 （英韓特許技術用語翻訳辞書）, http://plus.kipris.or.kr/portal/data/main/DBII_000000000000022/view.do?menuNo=210103&kppBCode=&kppMCode=&kppSCode=&subTab=&entYn=&clasKeyword=（最終検索日:2016 年 6 月 10 日）

・中韓対訳辞書[314]（2006 年）

ETRI（韓国電子通信研究院）が、中国語の Web 文書で頻度が高い単語を見出し語にして作成したデータである。

語数は 22 万語である。有料で研究、商用目的で利用できる。

・中韓新聞・ニュース用語対訳辞書[315]（2007 年）

ETRI が、情報検索用中韓新聞・ニュース分野で作成した品詞タグ付きの用語辞書である。

語数は 12.5 万語である。有料で研究、商用目的で利用できる。

・英韓・韓英科学技術分野専門辞書[316]（2007 年）

ETRI が、自動翻訳用専門辞書として作成したデータである。

語数は 96 万語である。有料で研究、商用目的で利用できる。

・中韓・韓中辞書[317]（1988 年）

SWRC 研究センターが、朝中辞書（朝鮮外国文図書出版社、中国民族出版社、1986 年）をベースにして作成した韓中・中韓辞書である。

語数は、中韓が 16 万語、韓中が 20 万語である。有料で研究、商用目的で利用できる。

・韓中対訳辞書[318]（2001 年）

SWRC 研究センターが、Ontology 等リソースをベースにして作成した対訳データである。

語数は５万語である。有料で研究、商用目的で利用できる

・韓中英の文パターン DB[319]（2000 年）

SWRC 研究センターが、用言を中心にその用言と他品詞との結合関係を模索して構築した中韓文パターンデータである。

文パターンで６万文である。有料で研究、商用目的で利用できる

---

314 https://etri.re.kr/tcenter/db/db_detail.jsp?historyId=23（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

315 https://etri.re.kr/tcenter/db/db_detail.jsp?historyId=24（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

316 https://etri.re.kr/tcenter/db/db_detail.jsp?historyId=25（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

317 Semantic Web Research Center, Electricdictionary7 Korean-Chinese dictionary, http://semanticweb.kaist.ac.kr/home/index.php/Electricdictionary7（最終検索日:2016 年 6 月 10 日）

318 Semantic Web Research Center, Electricdictionary9 Korean-Chinese Vocabulary, http://semanticweb.kaist.ac.kr/home/index.php/Electricdictionary9（最終検索日:2016 年 6 月 10 日）

319 Semantic Web Research Center, Electricdictionary10 Chinese-Korean sentence pattern, http://semanticweb.kaist.ac.kr/home/index.php/Electricdictionary10（最終検索日:2016 年 6 月 10 日）

(3) その他の辞書

・技術用語シソーラスデータ[320]

KIPRIS が、知的財産の技術用語との関連単語と単語の間の相関関係を数値化して作成した単語集である。

技術用語の語数は、146 万語である（2015 年 11 月 25 日更新)。無料であるが、商用目的の場合は事前協議及び許諾が必要である。

・公共ポータル公開辞書[321]

公共ポータルの OPEN API で公開された、以下の①～⑨の辞書からなる用例辞書である。

利用データの著作権を表示することで無料で商用・研究目的に利用できる。

① 観光用語の外国語用例辞書
② 電力関連用語辞典
③ 韓国文化財用語辞書
④ 関税・貿易関連用語辞書
⑤ 韓国歴史用語・シソーラス検索辞書
⑥ 済州のことわざ
⑦ 南北韓言語辞書
⑧ 国防科学技術用語情報
⑨ 軍事用語辞書

(4) 情報付けコーパス

・標準国語大辞典[322]（1999 年）

標準国語大辞典は標準語規定、ハングル正書法などの語文規定を遵守して国立国語院が 1999 年初版を発行した韓国語辞書である。データは標準語、方言、北韓語で分かれている。

語数は、51 万語である（2008 年 10 月更新)。無料の Web サービスが利用可能だが、商用目的では別途商談が必要である。

320 KIPRIS, 시 소 러 스 （シソーラス）, http://plus.kipris.or.kr/portal/data/service/DBII_000000000000021/view.do?pageIndex=12&menuNo=200101&kppBCode=&kppMCode=&kppSCode=&subTab=SC003&entYn=&clasKeyword=（最終検索日:2016 年 6 月 10 日）

321 한 국 정 보 화 진 흥 원 （NIA）, 공 공 데 이 터 포 털 , https://www.data.go.kr/main.jsp#/L21haW4=（最終検索日:2016 年 6 月 10 日）

322 국 립 국 어 원 , 표 준 국 어 대 사 전 , http://stdweb2.korean.go.kr/guide/entry.jsp（最終検索日:2016 年 6 月 10 日）

・韓国語情報付きコーパス[323]（2000 年）

国立国語院が、韓国語研究基盤を確保するために作成した情報付き韓国語コーパスで現代文語と現代口語がある。配布形態はソース、形態素解析、形態意味解析、構文解析である。ホームページではコーパス検索システムとコーパス統計情報を提供している。

データ量は、文語が 198 万 9,593 文（形態素解析 1013 万 344 語）、口語が 21 万 6,718 文（形態素解析 80 万 5,606 語）である。Creative Commons ライセンスで商用目的では利用できない。

その他、同サイトでは北朝鮮語と歴史的な形態素解析付きのコーパスも提供している。

(5) 韓国語ツリーバンク

・Korean Treebank Annotations Version 2.0[324]（2006 年）

ペンシルベニア州立大学の研究が開発した、Korean English Treebank Annotations（2002 年）の拡張版である。使用料金[325]は、Linguistic Data Consortium 会員の場合は無料、非会員の場合は＄500 である。

・Sejong Korean Treebank（2003 年）

韓国国立国語院が、世宗コーパス[326]をベースにして作成した 15 万文の構文解析情報付きの言語リソースである。

Creative Commons ライセンスで商用目的では利用できない。

・Korean dependency Treebank[327]（2005 年）

ETRI が Sejong,KAIST,ETRI コーパスから 20 語節以上の複文に形態情報、依存情報、構文関係情報を付けたデータである。

構築量は構文構造付き 10 万文である。有料で研究、商用目的で利用できる。

(6) 韓国語述語項付きツリーバンク

・Korean Propbank[328]（2006 年）

Korean English Treebank Annotations と Korean Treebank Version 2.0 における 13 万単語に、意味情報を付加したデータである。

---

323 국립국어원언어정보나눔터, http://ithub.korean.go.kr/user/guide/corpus/guide1.do（最終検索日:2016 年 6 月 10 日）

324 Chung-hye Han, Na-Rae Han "Part of Speech Tagging Guidelines for Penn Korean Treebank" IRCS Technical Report. University of Pennsylvania

325 Linguistic Data Consortium（LDC）, Korean Treebank Annotations Version 2.0, https://catalog.ldc.upenn.edu/LDC2006T09（最終検索日:2016 年 6 月 10 日）

326 https://ithub.korean.go.kr/user/main.do

327 https://etri.re.kr/tcenter/db/db_detail.jsp?historyId=22（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

328 Linguistic Data Consortium（LDC）, Korean Propbank, https://catalog.ldc.upenn.edu/LDC2006T03（最終検索日:2016 年 6 月 10 日）

(7) 韓国語文法規則リソース

・韓国語文規程[329]

国立国語院作成したもので、以下の①～⑤のデータからなる。

Creative Commons ライセンスで商用目的では利用できない。

① 標準語の規定（文部教育部告示第 88-2 号、1988.1.19 制定）[330]
② 追加標準語のリスト（最終：2015.12.14 公表-国語審議会の議決事項）
③ ハングル正書法（文化体育観光部告示第 2014-39 号、2014.12.5 一部改訂）
④ 外来語表記法（文化体育観光部告示第 2014-43 号、2014.12.5 一部改訂）
⑤ 国語のローマ字表記法（文化体育観光部告示第 2014-42 号、2014.12.5 一部改訂）

(8) 韓国語文法辞書

・外国人のための韓国語文法辞書[331]（2003 年）

国立国語院が 2000 年から実施した標準文法整備事業において作成された辞書である。この事業は、韓国語学習者、韓国語教師、韓国語教材編纂者が標準として利用できるようにするため、外国語としての韓国語文法を体系的に記述したものである。

2003 年における文法辞書の１次見出し語は 6,000 個。

「韓国語文法１，２」書籍で出版。

・標準国語文法開発[332]（2015 年）

国立国語院が、国語学会と国語教育学会及び一般言語生活の基準・参考とできるように、標準国語文法を開発することを目的にした研究である。

(9) 意味関係辞書

・CoreNet[333]（2002-2003 年）

SWRC 研究センターが、KORTERM 大量形態素解析コーパス[334]をベースに、単語の意味関係についての情報をまとめたデータである。

研究目的では無料で利用できるが、商用目的では別途有料の契約が必要である。

---

329 국립국어원, 한국 어문 규정 (2015년 현재)（韓国語文規定（2015 年現在））, http://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=485&pageIndex=1（最終検索日:2016 年 6 月 10 日）

330 국립국어원, 한글 맞춤법 （ハングル正書法）, http://www.korean.go.kr/front/page/pageView.do?page_id=P000060&mn_id=30（最終検索日:2016 年 6 月 10 日）

331 Eun-gyu Choi “外国語としての韓国語の文法研究”国立国語院、国語教育研究第１６集

332 Hyun-kyung Yoo “2015 年標準国語文法開発” 国立国語院研究課題最終報告書

333 Semantic Web Research Center, Core Multilingual Semantic Word Net, http://semanticweb.kaist.ac.kr/org/bora/CoreNet_Project/index.html（最終検索日:2016 年 6 月 10 日）

334 Semantic Web Research Center, KORTERM, http://semanticweb.kaist.ac.kr/research/korterm/korean/term.php（最終検索日:2016 年 6 月 10 日）

① 韓国語全体（韓中日対訳）：2,937 個の概念（４万 644 語彙）
② 多言語概念体系（韓中日）：2,937 個の概念名
③ 韓国語名詞：21,401 語彙（５万 1,607 意味）
④ 韓国語動詞：1,758 語彙（5,290 意味）
⑤ 韓国語形容詞：813 語彙（2,801 意味）
⑥ 韓国語動詞格フレーム（韓日対訳）：406 語彙（957 意味）
⑦ 韓国語形容詞格フレーム：759 語彙（1,109 意味）
⑧ 中国語全体（中韓対訳）：2,937 個の概念（２万 1,015 語彙）
⑨ 中国語名詞（中韓対訳）：２万 647 語彙
⑩ 中国語動詞格フレーム（中韓対訳）：288 語彙
⑪ 中国語形容詞格フレーム（中韓対訳）：80 語彙

・KorLex[335]（2004 年）

釜山大学校の人工知能研究室と韓国語情報処理研究室（KLPL）が、PWN（Princeton WordNet）を参照モデルにして 2004 年から構築した韓国語の語彙意味データである。

名詞、動詞、形容詞、副詞及び分類詞で構成されており、約 13 万語の意味セットと約 15 万語の語彙が含まれている。現在は 2.0 バージョンである。

研究目的の場合、KorLex 1.0 は無料で利用ができるが、その他のバージョンは有料である。商用目的の場合は、別途協議が必要である。

6.5.3 韓国語処理のための要素技術

(1) 単語分割技術

・KLT2000[336]（C++）

国民大学自然言語情報検索研究室が開発した韓国語形態素解析ツールであり、旧バージョン（HAM）の分割精度は 99.46%、スピードは約 4,500～5,000 単語/秒（Pentium PC/200MHz,Win95 環境下）である。

関連モジュールにとして以下がある。

① ハングル正書法検査及び較正機能
② 情報検索自動 INDEX 機能
③ ハングルの複合名詞分解機能
④ ハングル文章の自動分かち書き

---

335 PNU AI Lab & KLPL, About Korean Wordnet, http://korlex.pusan.ac.kr（最終検索日:2016 年 6 月 10 日）

336 강승식 교수：한글공학，정보검색 연구소（Dr. Kang's Korean Language Processing and Information Retrieval Laboratory），한국어 형태소 분석기와 한국어 분석 모듈（韓国語形態素解析と韓国語の分析モジュール）, http://nlp.kookmin.ac.kr/HAM/kor/index.html（最終検索日:2016 年 6 月 10 日）

研究目的であれば無料で利用できるが、商用目的では利用できない。

・HanNanum[337]（C/Java）

SWRC 研究センターが 1990 年に開発した韓国語形態素解析ライブラリであり、現在は既存の C から Java バージョンがリリースされている。

Plugin コンポネントアーキテクチャで柔軟性・拡張性が特徴。

KIST の品詞タグセットを利用している。

GPLv3 ライセンスに従って、研究・商用目的で利用可能。

・KKMA[338]（Java）

ソウル大学 IDS（Intelligent Data Systems）研究室が、自然言語処理のモジュール及びデータ構築の Project で開発した韓国語形態素解析ツールである。

研究目的は GPL2.0 に従って利用できるが、商用目的では別途協議が必要。

・KOMORAN 2.0[339]（Java）

SHINEWARE 自然言語処理研究所が開発した、確率基盤韓国語形態素解析器である。

空白が含まれている固有名詞の解析において、高い正確性が特徴である。

<速度評価>

（2.3GHz Intel Core i5、8GB 1600MHz DDR3）

３万語節解析（word/sec）

6,000 文解析（line/sec）

0.3MB 解析（MB/sec）

<正解率>

語節：91.37%（記号、数字含み）

品詞：94.60%（記号、数字含み）

現在は Ver.2.4 で Apache 2.0 ライセンスに従って、研究目的及び商用目的で無料で利用できる。

337 Semantic Web Research Center, HanNanum, http://semanticweb.kaist.ac.kr/home/index.php/HanNanum（最終検索日:2016 年 6 月 10 日）

338 서울대학교 IDS（Intelligent Data Systems） 연구실, 꼬꼬마 프로젝트!, http://kkma.snu.ac.kr（最終検索日:2016 年 6 月 10 日）

339 http://www.shineware.co.kr/?page_id=835（最終検索日:2016 年 6 月 10 日）（2016 年 8 月 2 日時点リンク切れ）

・Twitter Korean Text[340]（Scala/Java）

Twitter が Big Data から韓国語処理及び INDEX を抽出する目的に開発した韓国語処理ツールである。

現在はテキスト正規化、形態素解析、stemming（語幹化）、語句抽出機能がある。

Apache 2.0 ライセンスに従って、研究目的及び商用目的で無料で利用できる。

・Lucene Korean Analyzer[341]（Java）

Open Source 検索ライブラリ Lucene の、韓国語処理用形態素解析ライブラリである。

Apache 2.0 ライセンスに従って、研究目的及び商用目的で無料で利用できる。

・MACH2.0[342]（C）

School of Information Technology Sungshin Women's University が開発した高速韓国語形態素解析ツールである。

研究目的は無料で利用できる。商用目的では別途ライセンスが必要である。

<速度評価>

（3.3GHz Intel Core i5）

230 万語節解析（word/sec）

17MB 解析（MB/sec）

<正解率>

語節：99.2%（記号、数字含み）

・MeCab KO[343]

オープンソース形態素解析エンジン MeCab を用いた現代韓国語の解析ツールである。

解析用辞書は「HanDic」[344]を利用している。

ライセンスは MeCab を継承している。

・韓国語アナライザーRhino[345] [346]

韓国政府によって構築された 1,200 万の韓国語コーパスをベースに形態素解析と品詞分析を行うオープンソースのソフト。27 万以上の語幹と 8.5 万以上の語尾変化を解析でき

340 Will Hohyon Ryu（유 호 현）, Korean tokenizer, https://github.com/twitter/twitter-korean-text（最終検索日:2016 年 6 月 10 日）

341 NAVER Corp, 루 씬 한 글 분 석 기 , http://cafe.naver.com/korlucene（最終検索日:2016 年 6 月 10 日）

342 Kwangseob Shim and Jaehyung Yang, "MACH : A Supersonic Korean Morphological Analyzer", Proceedings of the 19th International Conference on Computational Linguistics

343 須賀井義教, コンピュータと朝鮮語のための覚え書き MeCab で韓国語, http://porocise.sakura.ne.jp/wiki/korean/mecab（最終検索日:2016 年 6 月 10 日）

344 須賀井義教, コンピュータと朝鮮語のための覚え書き HanDic の概要, http://porocise.sakura.ne.jp/wiki/korean/mecab/summary（最終検索日:2016 年 6 月 10 日）

345 http://lingua1972.blogspot.jp/2014/11/korean-morphological-analyzer-rhino.html

346 https://sourceforge.net/projects/koreananalyzer/

る。

(2) 品詞付け技術

・KLT2000（C++）

前記「単語分割技術」を参照のこと。

・HanNanum（C/Java）

前記「単語分割技術」を参照のこと。

・KKMA（Java）

前記「単語分割技術」を参照のこと。

・Khann2

コーパス制作における作業時間・費用を減らすために開発した品詞付けツールである。

・韓国語アナライザーRhino

前記「単語分割技術」を参照のこと。

(3) 構文解析技術

・KKMA（Java）

ソウル大学 IDS（Intelligent Data Systems）研究室が、自然言語処理のモジュール及びデータ構築の Project で開発した韓国語構文解析器ツールである。

研究目的は GPL 2.0 に従って利用できるが、商用目的では別途協議が必要。

・KLT2000[347]（C++）

国民大学校自然言語情報検索研究室が開発した韓国語構文解析器ツールである。

・KoreanParser[348]

SWRC 研究センターが自然言語処理のために構築した韓国語構文解析器である。

347 강승식 교수：한글공학，정보검색 연구소（Dr. Kang's Korean Language Processing and Information Retrieval Laboratory ），강승식 교수：한글공학，정보검색 연구소，http://nlp.kookmin.ac.kr

348 Semantic Web Research Center, KoreanParser, http://semanticweb.kaist.ac.kr/home/index.php/KoreanParser（最終検索日:2016 年 6 月 10 日）

・NLP-HUB[349]

SWRC 研究センターが国語情報処理システム競進大会で開発した構文解析及び形態素解析ツールである。

(4) その他の技術

・KoNLPy[350]

韓国語の自然言語処理をする Python ライブラリである。

GPLv3 ライセンスに従って利用できる。

・韓国語スペル・文法検査[351]

釜山大学校の人工知能研究室と NaraInfo 株式会社が 2001 年開発した韓国語のスペルと文法を検査するツールである。

・GrammE 1.0[352]

School of Information Technology Sungshin Women's University が開発した構文解析器の開発ツールである。

別のコーディングなしで、文法ルール開発だけで実際のシステムで使えるレベルの構文解析器開発が可能である。

研究目的は無料で利用できる。商用目的では別途ライセンスが必要である。

6.6 英語

6.6.1 英語処理のための言語リソース

(1) 英語と外国語間の対訳コーパス

特許データについては、パテントファミリーの公開特許公報から作成された対訳コーパスがある。

・ALAGIN の JPO・NICT 英日対訳コーパス

日本国特許庁と NICT の共同研究で作成され、高度言語情報融合フォーラム（ALAGIN）[353]から入手できる。件数は約３億 4,795 万文対で、利用条件は研究目的である。

---

349 Semantic Web Research Center, NLP HUB, http://semanticweb.kaist.ac.kr/home/index.php/NLP_HUB（最終検索日:2016 年 6 月 10 日）

350 KoNLPy: Korean NLP in Python, http://konlpy.readthedocs.io/en/v0.4.4/（最終検索日:2016 年 6 月 10 日）

351（주）나라인포테크, 한국어 맞춤법/문법 검사기, http://speller.cs.pusan.ac.kr/PnuSpellerISAPI_201602/（最終検索日:2016 年 6 月 10 日）

352 성신여자대학교, GrammE 1.0 - A Tool for Developing Your Own Syntactic Parser, http://cs.sungshin.ac.kr/~shim/demo/ge10.html（最終検索日:2016 年 6 月 10 日）

353 https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html

・日米特許全文対訳コーパス

日本の国立情報学研究所（NII）が研究目的で配布する日英対訳コーパス[354]である。データ範囲は 1993 年～2005 年、件数は約 318 万文対で、利用条件は研究目的である。

・米中特許全文対訳コーパス

日本の国立情報学研究所（NII）が研究目的で配布する英中対訳コーパス[355]である。データの範囲は 1993 年～2005 年、配布データは約 100 万文対、利用条件は研究目的限定である。

・韓国特許の英語抄録データ[356]（1979 年）

KIPRIS（特許庁）が韓国特許を英語文で加工して、海外に配布出来るように作成したデータである。

月単位でデータを更新し、2016 年 3 月 24 日更新で 199 万件である。有料で研究、商用目的で利用できる。データ利用料は、本年度が 282.6 万 KRW、過去分が 1,554.3 万 KRW である。

特許以外の主な対訳コーパスとして以下がある。

・Canadian Hansard Corpus[357] [358]

カナダ議会の議事録をもとに作成された英仏対訳コーパスである。1995 年に Linguistic Data Consortium（LDC）からリリースされた約 290 万文の対訳[357]の他に、作成者の異なるバージョンが存在する[358]。商用利用の可否は、データの提供者ごとに異なる。

・Europarl[359]

Moses の開発者である Philipp Koehn により構築された多言語対訳コーパスである。欧州議会の会議録に収録されている 21 の言語のテキストからアラインメントプログラムを用いて自動作成された。2012 年 5 月にリリースされた最新版の V7 では、例えば、独英は 192 万文、仏英は 201 万文の対訳が収録されている。商用利用可能である。

354 http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-ja-PatentMT.html
355 http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-ja-PatentMT.html
356 KIPRIS, BULK DATA 한국특허영문초록 （KPA）（韓国特許英文抄録）, http://plus.kipris.or.kr/portal/data/service/DBII_000000000000024/view.do?pageIndex=3&menuNo=200101&kppBCode=&kppMCode=&kppSCode=&subTab=SC003&entYn=&clasKeyword=（最終検索日:2016 年 6 月 10 日）
357 Hansard French/English, https://catalog.ldc.upenn.edu/LDC95T20（最終検索日:2016 年 6 月 30 日）
358 Aligned Hansards of the 36th Parliament of Canada, http://www.isi.edu/natural-language/download/hansard/（最終検索日:2016 年 6 月 30 日）
359 Europarl. http://www.statmt.org/europarl/（最終検索日:2016 年 6 月 30 日）

・Open Parallel Corpus（OPUS）[360]

OPUS は、インターネットから収集されたテキストをもとに作成した対訳コーパスである。映画やドラマのサブタイトルなどを利用して作成されている。ライセンス条件は不明である。

(2) 英語平文コーパス／注釈付きコーパス

学術研究用に収集されたコーパスの歴史は英語に関するものが最も歴史があり、特許以外の英語平文コーパスについても、他の言語に比べ様々なものが収集されている。すでに構築から時間が経過しており、最新状況の把握を目的とした調査の対象として必ずしも適当ではないものも含まれるかもしれないが、主なコーパスについてはある程度リストアップしておくことにする。

・Brown Corpus[361]

米ブラウン大学で 1960 年台に収集された最も初期のコーパスである。当時のアメリカ英語を収集したもので、規模は 100 万語。当初は平文コーパスであったが、後に品詞タグが付与されたバージョンがリリースされている。非営利目的での利用しかできない。

・Lancaster-Oslo-Bergen（LOB）Corpus[362]

ランカスター大学、オスロ大学、ベルゲン大学によって作成されたイギリス英語のコーパスである。上記の Brown Corpus と同時期に作成され、規模も 100 万語と同規模である。品詞タグが付与されている。

・British National Corpus（BNC）[363]

Oxford University Press や Longman Group、ランカスター大学などのコンソーシアムによって開発されたイギリス英語の均衡コーパスである。テキストの収集は 1990 年代前半に行われ、コーパスの規模は単語量で１億語である。品詞のタグが付与されている。商用利用可能である。

360 Jörg Tiedemann: Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation（LREC'2012）

361 BROWN CORPUS MANUAL. http://www.hit.uib.no/icame/brown/bcm.html（最終検索日:2016 年 6 月 30 日）

362 THE LANCASTER-OSLO/BERGEN CORPUS OF BRITISH ENGLISH. http://www.hd.uib.no/icame/lob/lob-dir.htm（最終検索日:2016 年 6 月 30 日）

363 British National Corpus, http://info.ox.ac.uk/bnc/（最終検索日:2016 年 6 月 30 日）

・Penn Treebank[364] [365] [366]

米ペンシルバニア大学によって開発された品詞タグ／構文タグが付与されたコーパスである。1995 年に LDC からリリースされた Treebank-2 には、人手でタグ付した 160 万語の Dow Jones News Service の記事データのほか Brown Corpus のテキストに構文タグを付与したデータ、100 万語の Wall Street Journal の記事などが含まれている[364]。またその後、変更されたタグセットを用いたデータが更改版としてリリースされている[365] [366]。商用利用可能性は LDC とのライセンス契約による。

・Lancaster Parsed Corpus[367]

LOB コーパスに収録されたテキストの中から約 1.2 万文に構文タグをつけたコーパスである。統計的構文解析を行った後に人手によりチェックしている。

・OA STM Corpus[368]

Elsevier が保有する科学技術論文誌のデータである。同社が保有する 1200 万件の記事の中で無料で読めるものが 60 万件あり、そのうちの 1.5 万件が Creative Commons ライセンスで利用できる Open Access データとなっている。アノテーションされているデータは少なく、品詞タグがついている記事が 110 件、構文タグがついている記事はわずか 10 件しか提供されていない。

(3) 英語単言語辞書

・New Oxford Dictionary of English, 2nd edition[369]

Oxford University Press による 17 万語を収録した英語辞書。ELRA から配布されている。商用利用可能である。

・COMLEX Syntax[370]

米ニューヨーク大学で開発された辞書で 3.8 万語について詳細な統語情報を収録している。LDC から配布されていたが、現在は LDC のカタログから削除されている。

364 Treebank-2. https://catalog.ldc.upenn.edu/LDC95T7（最終検索日:2016 年 6 月 30 日）
365 Treebank-3. https://catalog.ldc.upenn.edu/LDC99T42（最終検索日:2016 年 6 月 30 日）
366 English News Text Treebank: Penn Treebank Revised. https://catalog.ldc.upenn.edu/LDC2015T13（最終検索日:2016 年 6 月 30 日）
367 Lancaster Parsed Corpus. http://clu.uni.no/icame/manuals/LPC/LPC.PDF（最終検索日:2016 年 6 月 30 日）
368 OA STM Corpus, http://elsevierlabs.github.io/OA-STM-Corpus/（最終検索日:2016 年 6 月 30 日）
369 ELRA-L0045:New Oxford Dictionary of English, 2nd Edition, http://catalog.elra.info/product_info.php?products_id=679（最終検索日:2016 年 6 月 30 日）
370 COMLEX Syntax, http://nlp.cs.nyu.edu/comlex/（最終検索日:2016 年 6 月 30 日）

・VALEX[371]

英ケンブリッジ大学で開発された、動詞を対象とした辞書である。約 6,400 語に関する下位範疇化フレーム（subcategorization frame.SCF）と頻度の情報を有し、下位範疇化は 163 種のタイプによって分類されている。研究目的での利用のみが可能である。

(4) シソーラス、概念辞書

・WordNet[372]

米プリンストン大学によって開発された概念辞書である。英語の名詞、動詞、形容詞、副詞を 11.7 万の synset と呼ぶ概念グループに分類し、それらの間の上位下位関係などが約 20 万件定義されている。この WordNet の開発を受けて各国で同様の概念辞書の開発が行われたが、WordNet 自体は 2006 年を最後にアップデートされていない。商用利用可能である。

・Oxford Paperback Thesaurus, 2nd edition[373]

33 万語を収録するシソーラスである。2.9 万語に対しては、コーパスから抽出した事例が付与されている。ELRA から配布されている。商用利用可能である。

6.6.2 英語処理のための要素技術

(1) 形態素解析

英語の形態素解析に関しては既に種々のプログラムがオープンソースとして公開されている。英語は空白で分かち書きされるため、品詞付与をしない場合は SMT の代表的オープンソースである Moses に同梱されている tokenizer を利用するだけで問題ない。

・Stanford Log-linear Part-Of-Speech Tagger[374] [375]

スタンフォード大学が開発した品詞付けツール。Bidirectional dependency network 等を用いた手法により Penn Treebank の WSJ のデータに対する実験で 97.24%の品詞付与精度が得られたことが報告されている[374]。最新の更新は 2015 年 12 月である。商用利用可能である。

---

371 VALEX - A Large Subcategorization Lexicon for English Verbs , http://www.cl.cam.ac.uk/~alk23/subcat/lexicon.html（最終検索日:2016 年 6 月 30 日）

372 WordNet, https://wordnet.princeton.edu/（最終検索日:2016 年 6 月 30 日）

373 ELRA-L0048:Oxford Paperback Thesaurus, 2nd edition, http://catalog.elra.info/product_info.php?products_id=682（最終検索日:2016 年 6 月 30 日）

374 Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

375 Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml（最終検索日:2016 年 6 月 30 日）

(2) 構文解析

主な構文解析（係り受け解析・依存解析を含む）のツールとして以下がある。

・Apple Pie Parser[376]

米ニューヨーク大学で開発された統計的チャートパーザである。ツールに同梱されている英語の文法は Penn Treebank から機械的に抽出されたものが用いられている。1997 年以降更新されていない。

・Berkley Parser[377]

米カリフォルニア大学バークレー校で開発された構文情報の注釈がついた Penn Treebank のデータを用いて学習した確率文脈自由文法（Probabilistic Context-Free Grammar,PCFG）を用いて解析を行う構文解析器。Java で実装されている。商用利用可能である。

・Stanford Parser[378]

スタンフォード大学で開発された語彙化された PCFG を用いた構文解析器。GNU GENERAL PUBLIC LICENSE の下で商用利用可能である。

・Ckylark[379] [380]

NAIST の小田悠介氏らによって構築された PCFG（確率的文脈自由文法）による構文解析器。従来の解析器に対し頑健性の向上を図っている。

・Egret[381]

Hui Zhang 氏らによって構築された PCFG による構文解析器であり、上で述べた Berkeley Parser を再実装したもの。曖昧性を圧縮した統語ベースの出力が可能である。Apache 2.0 並びに LGPL 3.0 のデュアルライセンスの下で商用利用可能である。

・Enju[382] [383]

東京大学の辻井研究室を中心に構築された HPSG による英語の構文解析器であり、述語項解析も行える。

---

376 Apple Pie Parser. http://nlp.cs.nyu.edu/app/（最終検索日:2016 年 6 月 30 日）

377 berkeleyparser, https://github.com/slavpetrov/berkeleyparser（最終検索日:2016 年 6 月 30 日）

378 The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml（最終検索日:2016 年 6 月 30 日）

379 Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda and Satoshi Nakamura : Ckylark: A More Robust PCFG-LA Parser, Proceedings of NAACL-HLT 2015, pages 41–45, 2015.

380 Ckylark HP : https://github.com/odashi/Ckylark（最終検索日:2016 年 6 月 30 日）

381 Egret HP : https://sites.google.com/site/zhangh1982/egret（最終検索日:2016 年 6 月 30 日）

382 Yusuke Miyao and Jun'ichi Tsujii : Feature Forest Models for Probabilistic HPSG Parsing, Computational Linguistics, Vol. 34, No. 1, Pages 35-80, 2008.

383 Enju HP : http://www.nactem.ac.uk/enju/index.ja.html（最終検索日:2016 年 6 月 30 日）

・MaltParser[384] [385]

スウェーデンの Växjö 大学と Uppsala 大学で開発された依存解析（係り受け解析）器の生成システムである。文法から解析器を生成する代わりに依存構造の注釈がついたコーパスを与えて、解析器を生成する。生成される解析器は決定的な解析を行う。当初は研究目的の利用に限定されていたが、現在は商用利用可能である。

・Apache OpenNLP[386]

Apache Software Foundation が提供するオープンソースの自然言語処理ツールキット。ツールの中に構文解析器が含まれており、英語などのモデルが同梱されている。

6.7 日本語

ここでは、日本語及び日本語と外国語間の言語リソースの現状を説明する。特許文書の機械翻訳に関連が深いものに限り、網羅的に説明するものではない。例えば、音声翻訳に関する言語リソースに関しては触れない。また、各リソースの利用条件の詳細については、参考文献を参照されたい。

6.7.1 日本語処理のための言語リソース

(1) 日本語と外国語間の特許対訳コーパス

特許データについては、パテントファミリーの公開特許公報から作成された対訳コーパスがある。作成者は特許庁、NICT、日本国特許庁と NICT の共同研究である。作成されたデータは高度言語情報融合フォーラム（ALAGIN）と特許庁から入手できる。データ範囲が少し古い国立情報学研究所（NII）が配布するデータもある。入手先と利用条件を一覧する。

384 Nivre, J., J. Hall and J. Nilsson: MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In Proceedings of the fifth international conference on Language Resources and Evaluation（LREC）, pp.2216-2219, 2006.

385 MaltParser http://maltparser.org/（最終検索日:2016 年 6 月 30 日）

386 Apache OpenNLP https://opennlp.apache.org/（最終検索日:2016 年 6 月 30 日）

表６．７．１-１　日本語と外国語間の特許対訳コーパス

1	日本語 ⇔英語	ALAGIN の JPO・NICT 英日対訳コーパス（約３億 4,795 万文対） 研究目的 https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html
2	日本語 ⇔英語	日米特許全文対訳コーパス（318 万 6,284 文対） 研究目的　NTCIR　データ範囲は 1993 年～2005 年 http://research.nii.ac.jp/ntcir/permission/ntcir-9/perm-ja-PatentMT.html
3	中国語 ⇔日本語	特許庁　中日対訳コーパス（約１億 7,318 万文対） 庁内利用・一部研究目的　データ範囲は 2005 年～2014 年 https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyo_dictionary.htm
4	中国語 ⇔日本語	ALAGIN の JPO 中日対訳コーパス（約１億 3,285 万文対） 研究目的　データ範囲は 2005 年～2013 年（上記に含まれる） https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html
5	韓国語 ⇔日本語	ALAGIN の JPO・NICT 韓日対訳コーパス（約 8,346 万文対） 研究目的 https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html

また、特許以外の主な対訳コーパスとして以下がある。

・アジア学術論文抜粋コーパス（ASPEC）[387] [388]

約 300 万対訳文からなる日英論文抄録コーパス（ASPEC-JE）と約 68 万対訳文からなる日中論文抜粋コーパス（ASPEC-JC）から成る大規模な論文対訳コーパス。国立研究開発法人科学技術振興機構（JST）と国立研究開発法人情報通信研究機構（NICT）が 2006 年から 2010 年にかけて構築した。研究目的で利用できる。

・日英新聞記事対応付けデータ（JENAAD）[389]

NICT が構築した読売新聞と The Daily Yomiuri から自動作成された日英対応付けコーパスであり、30 万の一対一文対応データと６万の一対多文対応データからなる。商用利用も可能である。

[387] ASPEC HP http://orchid.kuee.kyoto-u.ac.jp/ASPEC/（最終検索日:2016 年 6 月 24 日）

[388] Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, Hitoshi Isahara : ASPEC: Asian Scientific Paper Excerpt Corpus, Proceedings of the Ninth International Conference on Language Resources and Evaluation, pages 2204-2208, 2016.

[389] Masao Utiyama and Hitoshi Isahara : Reliable Measures for Aligning Japanese-English News Articles and Sentences, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 72--79, 2003.

(2) 日本語平文コーパス[390]

電子化されている特許リソースは 2016 年現在下記のとおりである。

・公開特許公報（1993 年～2003 年、405 万件 SGML）
・公開特許公報（2003 年～2015 年、405 万件 XML）

また、特許以外の主な日本語平文コーパスとして以下のようなものがある。

・現代日本語書き言葉均衡コーパス（BCCWJ）[391]

現代日本語の書き言葉の全体像を把握するために国立国語研究所が構築したコーパスであり、主として 1986 年から 2005 年に刊行された書籍全般、雑誌全般、新聞、白書、ブログ、ネット掲示板、教科書、法律などのジャンルにまたがって１億 430 万語のデータを格納してある。商業目的での利用も検討可能である。

・国語研日本語ウエブコーパス [391]

国立国語研究所が構築中の 100 億語を超える規模の現代日本語コーパスでウェブ（WWW）上の日本語テキストを利用して構築している（開発期間 2011 年度～2016 年度)。

・各種新聞データ

各種新聞データであり、日外アソシエーツ株式会社から配布されている[392]。学術研究用で利用可能である。

毎日新聞（1991 年～2015 年、9～11 万記事／年）
朝日新聞（1988 年～2015 年、14 万記事前後／年）
読売新聞（1987 年～2015 年、8～38 万記事／年）

(3) 日本語注釈付きコーパス[393]

・京都大学テキストコーパス[394]

毎日新聞の記事に対して Juman/KNP で係り受け解析をほどこし人手で修正したコーパス（約４万文)。京都大学の黒橋・河原研究室が作成した。本パッケージに含まれるのは形態素・構文・関係の付加情報だけで、もとの毎日新聞データは含まれていない。コーパス本来の形に変換するには毎日新聞 1995 年版 CD-ROM が必要である。

390 構文解析・意味解析などの言語分析を加えていないテキストのみのコーパスを平文コーパスと呼ぶ。
391 国立国語研究所 HP http://pj.ninjal.ac.jp/corpus_center/（最終検索日:2016 年 6 月 24 日）
392 日外アソシエーツ HP http://www.nichigai.co.jp/sales/corpus.html（最終検索日:2016 年 6 月 24 日）
393 単純なテキストだけではなく、構文解析や意味解析など各種の言語解析の結果を注釈として付与したコーパス。
394 京都大学黒橋・河原研究室 HP http://nlp.ist.i.kyoto-u.ac.jp/（最終検索日:2016 年 6 月 24 日）

・BCCWJ:述語項構造と照応タグつきコーパス[395] [396]

「日本語コーパス:代表性を有する大規模日本語書き言葉コーパスの構築」ツール班が作成した表題の注釈付きコーパスであり、2011 年現在、約２万文から構成されている。コーパスデータの復元には日本語書き言葉均衡コーパス（BCCWJ）の DVD 版のデータが必要となる。タグ本体は修正 BSD ライセンスにて配布されている。

(4) 特許対訳辞書

特許用語の対訳辞書として日本国特許庁が作成した辞書が特許庁と高度言語情報融合フォーラム（ALAGIN）から入手できる。入手先と利用条件を表６．７．１-２に一覧する。

表６．７．１-２　特許対訳辞書一覧

1	日英辞書	特許庁　日英機械翻訳辞書データ（約 9.6 万語） 公開・商用利用可 https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyo_dictionary_dl.htm
2	中日辞書	特許庁　中日機械翻訳辞書（約 226 万語） 庁内利用 https://www.jpo.go.jp/shiryou/toushin/chousa/tokkyo_dictionary.htm
3	中日辞書	ALAGIN の JPO 中日対訳辞書（約 220 万語：上記に含まれる） 研究目的 https://alaginrc.nict.go.jp/resources/jpo-info/jpo-outline.html

また、特許以外の主な対訳辞書として以下がある。

・EDR 対訳辞書[397]

日本電子化辞書研究所（EDR）が 1986 年から 1994 年に作成した辞書で、英日対訳辞書が基本語 16 万語、日英対訳辞書が基本語 23 万語、日中対訳辞書が基本語 23 万語を収録している。商用利用可能である。

[395] 小町守,飯田龍:BCCWJ に対する述語項構造と照応関係のアノテーション,日本語コーパス平成 22 年度公開ワークショップ,pages 325-330. March 2011.

[396] 奈良先端科学技術大学院大学自然言語処理研究室 HP http://cl.naist.jp/（最終検索日:2016 年 6 月 24 日）

[397] EDR HP https://www2.nict.go.jp/out-promotion/techtransfer/EDR/J_index.html（最終検索日:2016 年 6 月 24 日）

・JMdict/EDICT[398]

Electronic Dictionary Research and Development Group が作成した辞書で、日本語を中心として、英独仏露日蘭などの諸語との対訳辞書である。2016 年現在で、日英部分は約 17 万エントリーを収録している。商用利用可能である。

・CICC 辞書[399]

（財）国際情報化協力センター（CICC）が 1995 年に作成した辞書で、マレーシア語、インドネシア語、中国語、タイ語の各言語の辞書である。辞書のエントリーには対応する英語も付与されている（中国語辞書を除く）。言語資源協会（GSK）から配布されている。語数は以下のとおりである。商用利用はできない。

マレーシア語　７万語

インドネシア語　５万語

中国語　５万語

タイ語　５万語

専門用語辞書（コンピュータ、電気、工学、及び関連分野）　2.7 万語（上記５カ国語に加えて日本語と英語の対訳も付与）

・中日専門用語辞書（CJ-TERM）[400]

株式会社日中韓辭典研究所が作成した理工学系の専門用語辞書であり、82 万語を含んでいる。商用利用可能である。

・鳥バンク[401]

日本語表現意味辞書等管理委員会が作成した日英対訳の意味類型パターン辞書であり 22.7 万パターン対からなる。研究目的で利用できる。

・日本語語彙体系（文型パターン部分）[402] [403]

NTT の池原悟氏らが著し、岩波書店から販売されている語彙体系に含まれる 1.4 万件の文型パターンである。3,000 種の意味分類を用いて日本語の文型を定義しており、そのすべてに英語文型が付与されている。研究利用が可能である。

398 JMdict/EDICT HP http://www.edrdg.org/jmdict/edict_doc.html（最終検索日:2016 年 6 月 24 日）

399 GSK HP http://www.gsk.or.jp/catalog/（最終検索日:2016 年 6 月 24 日）

400 日中韓辭典研究所 HP http://www.cjk.org/cjk/samples/cjtermj.htm（最終検索日:2016 年 6 月 24 日）

401 鳥バンク http://unicorn.ike.tottori-u.ac.jp/toribank/（最終検索日:2016 年 6 月 24 日）

402 日本語語彙大系 CD-ROM 版 http://www.kecl.ntt.co.jp/icl/lirg/resources/GoiTaikei/（最終検索日:2016 年 6 月 24 日）

403 池原悟,宮崎正弘,白井諭,横尾昭男,中岩浩巳,小倉健太郎,大山芳史,林良彦: 日本語語彙大系 CD-ROM 版. 岩波書店, 1999.

(5) 日本語辞書

① 単語辞書

・ipadic[404]

奈良先端科学技術大学院大学（NAIST）が開発した日本語形態素解析システム ChaSen 用の辞書であり、情報処理振興事業協会（IPA）が設定した IPA 品詞体系に基づいている。23.9 万語から構成されている 2007 年版が公開されている。

・UniDic[405]

奈良先端科学技術大学院大学（NAIST）が開発した日本語形態素解析システム MeCab 用の辞書であり、国立国語研究所が規定した「短単位」という揺れのない斉一な単位で設計されている。75.6 万語から構成されている 2013 年版が公開されている。

・EDR 日本語辞書 [397]

日本電子化辞書研究所（EDR）が 1986 年から 1994 年に作成した辞書で、基本語 27 万語を収録した辞書と情報処理関係の専門用語 11 万語を収録している。商用利用可能である。

②係り受け辞書

・EDR 日本語共起辞書 [397]

日本電子化辞書研究所（EDR）が 1986 年から 1994 年に作成した辞書で、係り受け関係にある 90 万句が収録されている。商用利用可能である。

・EDR 英語共起辞書 [397]

日本電子化辞書研究所（EDR）が 1986 年から 1994 年に作成した辞書で、係り受け関係にある 46 万句が収録されている。商用利用可能である。

・日本語係り受けデータベース

NICT が構築し ALAGIN[406]から配布されている係り受けデータベースで、大量の日本語の Web 文書（約６億ページ、430 億文、クロール時期は 2007 年５月 19 日から 11 月 13 日）のデータを Juman/KNP で係り受け解析した結果から、語句と語句の係り受けを抽出し、ある程度のノイズデータを取り除いた上で、係り受けとその頻度を収録したもので、約 46 億種類の係り受けが含まれている。

404 IPAdic legacy https://osdn.jp/projects/ipadic/（最終検索日:2016 年 6 月 24 日）
405 UniDic https://osdn.jp/projects/unidic/（最終検索日:2016 年 6 月 24 日）
406 ALAGIN HP https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html（最終検索日:2016 年 6 月 24 日）

利用には ALAGIN に加入するとともに、NICT とライセンス契約が必要である。

③格フレーム辞書

・京都大学格フレーム [394]

京都大学の黒橋・河原研究室が Web テキストから自動構築した大規模格フレーム。格フレームとは、用言とそれに関係する名詞を用言の用法ごとに整理したものである。この格フレームは、Web 上の約 16 億文の日本語テキストから自動的に構築しており、約４万用言からなる。言語資源協会（GSK）から配布されている。

・京都大学名詞格フレーム [394]

京都大学の黒橋・河原研究室が Web テキストから自動構築した大規模名詞格フレーム。名詞格フレームとは、名詞とその意味を解釈する上で必須となる要素を名詞の語義ごとに整理したものである。例えば、「値段」という名詞に対しては『品物』、「レバー」という名詞に対しては「機械を操作するための棒」、「動物の肝臓」という語義ごとにそれぞれ『機械』、『動物』が収集されている。この名詞格フレームは、Web 上の約 16 億文の日本語テキストから自動的に構築しており、約 16 万名詞からなる。

(6) シソーラス

・日本語語彙体系（語彙部分）[402] [403]

NTT の池原悟氏らが著し、岩波書店から販売されている語彙の体系である。30 万語の収録語が 3,000 種類の意味分類を用いて定義されており、最大規模の日本語シソーラスとなっている。研究利用が可能である。

・分類語彙表[407]

国立国語研究所が 1964 年に作成し、2004 年に増補改訂した「語を意味によって分類・整理したシソーラス（類義語集）」である。10 万件のレコードから構成されており、分類項目内の意味区切りを示すレコードを含んでいる。学術研究用に利用できる。

・EDR 概念辞書 [397]

日本電子化辞書研究所（EDR）が 1986 年から 1994 年に作成した辞書で、基本語 41 万概念を収録したシソーラス辞書である。商用利用可能である。

[407] 分類語彙表-増補改訂版データベース　https://www.ninjal.ac.jp/archives/goihyo/（最終検索日:2016 年 6 月 24 日）

・日本語 WordNet[408] [409]

NICT の Francis Bond 氏（現 Nanyang Technological University）らが作成したデータであり、英語の WordNet との平行性がある。約 9.3 万単語、5.7 万概念、15.8 万語義などが記述されている。

・上位語階層データ[410]

NICT が構築し ALAGIN から配布されているデータで、Wikipedia（2007-03-28 版）から得られた上位下位関係に現れた上位語、約 6.9 万名詞句を階層化して、階層を構成する名詞句のすべてに、その指示対象が十分に特定されるかどうかのタグ付けを行ったものである。

(7) N-gram データ・共起辞書

・Web 日本語 N グラム第 1 版[411]

Google が日本語 Web ページから抽出した N-gram データであり、出現頻度 20 回以上の 1 から 7-gram を収録している。異なり数は 1-gram で 256 万、7-garm で 5 億 7,020 万である。2007 年に GSK から公開された [399]。教育・研究目的で利用できる。

・単語共起頻度データベース[412]

NICT が作成し ALAGIN から配布されているデータベースであり、約 1 億の Web ページを用いて、様々な条件で、２つの単語が共起する回数を計算して、各単語（最大約 100 万語）について３種類の共起スコア（共起頻度、Dice 係数、ディスカウンティング相互情報量）の高い順に、最大 100 単語を列挙したものである（他に全データの配布もある）。

6.7.2 日本語処理のための要素技術

言語処理ツールの現状について日本語処理を中心にして説明する。また、言語処理は二つの言語を対で扱うことが多いため、日本の開発者による英語処理・中国語処理のツールもあげた。なお、日本発の言語処理ツールは非常に多いため、特許文書の機械翻訳に関連が深いものと学会発表・書籍等でよく使われるツールを取り上げた。

---

408 日本語 WordNet HP : http://compling.hss.ntu.edu.sg/wnja/（最終検索日:2016 年 6 月 24 日）

409 Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki : Enhancing the Japanese WordNet in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009.

410 https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html
（A-4）上位語階層データ

411 大規模日本語 n-gram データの公開　http://googlejapan.blogspot.jp/2007/11/n-gram.html（最終検索日:2016 年 6 月 24 日）

412 GSK HP http://www.gsk.or.jp/catalog/（最終検索日:2016 年 6 月 24 日）

(1) 形態素解析

主な形態素解析ツールとして以下がある。

・JUMAN[413]

京都大学の黒橋・河原研究室が構築した形態素解析ツールであり、使用者によって文法の定義，単語間の接続関係の定義などを容易に変更できるように配慮してある。

・ChaSen[414] [415]

奈良先端科学技術大学院大学の松本裕治研究室が構築した形態素解析ツールであり、JUMAN から分岐し、Hidden Markov Model（HMM）をベースにしたシステムである。商用利用可能である。

・MeCab[416]

京都大学情報学研究科-日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンである。言語、辞書、コーパスに依存しない汎用的な設計を基本方針としており、パラメータの推定に Conditional Random Fields（CRF）を用いている。

・KyTea[417] [418]

Graham Neubig 氏らが構築した、日本語など、単語（又は形態素）分割を必要とする言語のための一般的なテキスト解析器である。線形 SVM（Support Vector Machine）やロジスティック回帰などを用いてそれぞれの分割点や読みを個別に推定するため、部分的にアノテーションされたデータを利用してモデルを学習することも可能である。

(2) 構文解析

主な構文解析（係り受け解析・依存解析を含む）のツールとして以下がある。

413 JUMAN HP : http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN（最終検索日:2016 年 6 月 24 日）
414 ChaSen -- 形態素解析器 HP : http://chasen-legacy.osdn.jp/（最終検索日:2016 年 6 月 24 日）
415 松本裕治:形態素解析システム「茶筌」, 情報処理学会誌, Vol.41, No.11, pages 1208-1214, 2000.
416 MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html（最終検索日:2016 年 6 月 24 日）
417 KyTea HP : http://www.phontron.com/kytea/index-ja.html（最終検索日:2016 年 6 月 24 日）
418 Graham Neubig, Yosuke Nakata, Shinsuke Mori : Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

・KNP[419] [420]

京都大学の黒橋・河原研究室が構築した日本語文の構文・格・照応解析を行うシステムである。形態素解析システム JUMAN の解析結果（形態素列）を入力とし、文節及び基本句間の係り受け関係、格関係、照応関係を出力する。係り受け関係、格関係及び照応関係は、Web から自動構築した大規模格フレームに基づく確率的モデルにより決定する。

・CaboCha[421] [422]

工藤拓氏らによって構築された Support Vector Machines（SVM）に基づく日本語係り受け解析器であり、生文はもちろん、形態素解析済みデータ、文節区切り済みデータ、部分的に係り関係が付与されたデータからの解析が可能である。処理系は商用利用可能であるが、付属のモデルは研究利用に限られる。

・Ckylark[423] [424]

小田悠介氏らによって構築された PCFG（Probabilistic Context Free Grammar,確率的文脈自由文法）による構文解析器であり、英語と日本語のモデルがある。

・Egret[425]

Hui Zhang 氏らによって構築された PCFG による構文解析器であり、曖昧性を圧縮した統語森での出力が可能である。英語、中国語に加えて日本語のモデルもある[426]。

・CNP[427] [428]

NICT によって構築され ALAGIN から配布されている中国語依存解析器である。中国語解析モデルも用意されている[429]。商用利用可能である。

---

419 KNP HP : http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP

420 笹野遼平,河原大輔,黒橋禎夫,奥村学:構文・述語項構造解析システム KNP の解析の流れと特徴,言語処理学会 第 19 回年次大会論文集, pages 110-113, 2013.

421 CaboCha/南瓜: Yet Another Japanese Dependency Structure Analyzer, http://taku910.github.io/cabocha/（最終検索日:2016 年 6 月 24 日）

422 工藤拓,松本裕治：チャンキングの段階適用による日本語係り受け解析、情報処理学会論文誌, Vol.43, No.6, pages 1834-1842, 2002.

423 Ckylark HP : https://github.com/odashi/Ckylark（最終検索日:2016 年 6 月 24 日）

424 Yusuke Oda, Graham Neubig, Sakriani Sakti, Tomoki Toda and Satoshi Nakamura : Ckylark: A More Robust PCFG-LA Parser, Proceedings of NAACL-HLT 2015, pages 41–45, 2015.

425 Egret HP : https://sites.google.com/site/zhangh1982/egret（最終検索日:2016 年 6 月 24 日）

426 Neubig / egret HP : https://github.com/neubig/egret（最終検索日:2016 年 6 月 24 日）

427 CNP - A ChiNese dependency Parser : https://alaginrc.nict.go.jp/cnp/index.html（最終検索日:2016 年 6 月 24 日）

428 Wenliang Chen, Jun'ichi Kazama, Kiyotaka Uchimoto, and Kentaro Torisawa : Improving Dependency Parsing with Subtrees from Auto-Parsed Data, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 570–579, 2009.

429 A Chinese Dependency Parser（CNP）用中国語解析モデル : https://alaginrc.nict.go.jp/resources/nict-resource/li-info/li-outline.html#C-2（最終検索日:2016 年 6 月 24 日）

・Enju[430] [431]

東京大学の辻井研究室を中心に構築された HPSG（Head-driven phrase structure grammar）による英語の構文解析器であり、述語項解析も行える。

(3) 述語項解析

・KNP[419]

3.2 で述べた KNP では述語項解析（格解析）が可能である。

・SynCha[432] [433]

飯田龍氏らが構築した述語項解析システムであり、3.2 で述べた CaboCha をベースにしている。

・Enju[430] [431]

東京大学の辻井研究室を中心に構築された HPSG による英語の構文解析器であり、述語項解析も行える。

430 Enju HP : http://www.nactem.ac.uk/enju/index.ja.html（最終検索日:2016 年 6 月 24 日）

431 Yusuke Miyao and Jun'ichi Tsujii : Feature Forest Models for Probabilistic HPSG Parsing, Computational Linguistics, Vol. 34, No. 1, Pages 35-80, 2008.

432 SynCha HP : https://sites.google.com/site/ryuiida/syncha（最終検索日:2016 年 6 月 24 日）

433 飯田龍,小町守,井之上直也,乾健太郎,松本裕治:述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から、自然言語処理, Vol.17, No.2, pages 25-50, 2010.