有限会社サイバープロ

本社住所	〒525-0011 滋賀県草津市片岡町292－5
URL	http://www.cyberpro.co.jp
展示名	**データベースアプローチに基づく特許機械翻訳システム**
お問合せ先	担当部署　営業部 TEL　077-568-1931 FAX　077-568-1931 E-mail　cyberoikeda@yahoo.co.jp

■アピールポイント

従来の機械翻訳アプローチである、ルールベース、統計ベース、事例ベースに対し、データベースアプローチは、完全対訳節と完全名詞句からなるデータベースの構築と、その応用としての機械翻訳という2段階アプローチをとっている点であり、節の構造を分解しないで、対訳節を対応させている点で特徴がある。これにより、従来自然でわかりやすい翻訳のためには、対訳文の構造を代えなければいけない文の翻訳品質を格段に向上させることができる。また、翻訳に対し、なぜこのような訳になっているかを示すアカウンタビリティ機能がついている点で特徴がある。

【産業日本語との関連】

制限言語としての産業日本語の確立のためには、語や文法の制約ではその記述能力の保証ができない。そのためには、文中での意味の単位である「節」と、「完全名詞句（節を含まないすべての名詞句）」の内、対象文書記述のためのものを抜出し、制約しなければいけない。更に、機械翻訳の品質を保証するためには、制約した節と名詞句に対する対訳を一意に決めておかなければいけない。この2つの制約を課した文の作成を支援したり、自由に書かれた文を同意の制約文に書き換えたり、制約文を翻訳したりしようとするのが本システムで、産業日本語成立のまさに中心課題を解決するものである。

【詳細】

文の節分解は、例えば次のように行う。

SO=このようにして車体を浮上させた場合には、摩擦駆動は行われず、磁気誘導による推進駆動、さらにはこの推進駆動にプロペラによる補助推進駆動を加えた推進駆動となる。	S0= In the case where the chassis is made to float in this manner, frictional drive is not provided, and propelling drive derived from magnetic induction, or auxiliary propelling drive using propellers is added.
S0=@連用形タ接続:た場合には、@受動態-未然形:ず、_、_。([S1],[S2],[S3]) S1=このようにして_を浮上させる([N4]) S2=[_が]摩擦駆動を行う([:人]) S3=_、さらには_を加えた_となる([N5],[N6],[N7]) N4=車体 N5=磁気誘導 N6=推進駆動 N7=プロペラによる補助推進駆動	S0=In the case where @Passive:, @Passive and @Passive:.([S1],[S2],[S3]) S1= the _ is made to float in this manner([N4]) S2=[_] provide frictional drive([person]) S3=[_] add _, or _ to _([N5],[N6],[N7]) N4= chassis N5= magnetic induction N6= propelling drive N7= auxiliary propelling drive using propellers

この文(関数列)分解の特長は、分解要素（関数）がすべて節（文を含む）か名詞句になっている

ことで、関数内の埋め込みには、活用、態、時制などの変換が伴っていることである。このようにしても、S1 のように使役形日本文が受動態英文の対応することや、S3 のように内部統語構造は異なるものになることや、N7 のように複合名詞句を分解せずに翻訳しないとうまく対応しないものがある。また、これらがルールベースの機械翻訳で誤訳になる1つの原因であった。

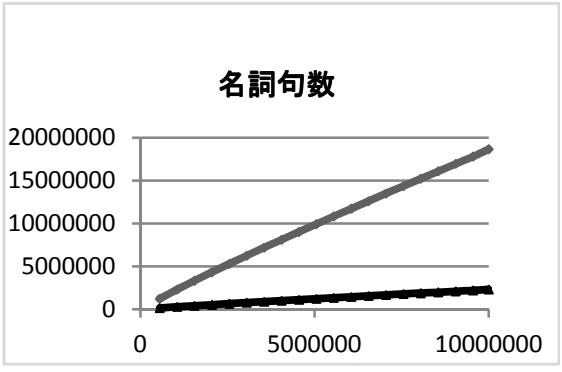
本システムは、これらの問題を解決するため、実在する特許翻訳文（NTCIR-10）から、すべての対訳節、対訳名詞句を自動的に切り出してデータベース化した。そして、その対訳対の内部構造での対応もすべてデータベースに格納した。すべての要素（ここでは、節と名詞句）をすべて具体的な表層レベルで管理するのが、データベースアプローチという所以である。従来のルールベースは、少ない要素（レキシコンと品詞並び規則）ですべての文の表層構造を理解しようとしてうまくいかなかったが、データベースアプローチは、品詞という中間構造物を使わず、具体的な表層レベルでの表現を基本にしている点が重要である。

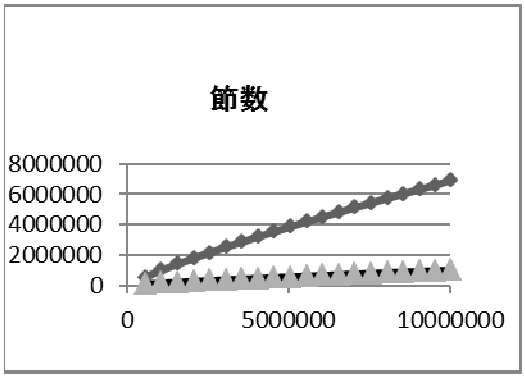
また、要素分解の段階で、従来共起(collocation)として扱われていた、動詞―目的語、副詞―動詞、形容詞―名詞などの従属関係を、すべて節や名詞の中に閉じ込めてしまった。例えば、上例の S2 のように、[摩擦駆動を行う]-[provide frictional drive]という対応は、分解して、[[摩擦駆動]-[frictional drive]と[_を行う([N])]-[provide _([N])]の2つの要素には分解できず、一体として扱わなければいけない。これも誤訳の一因であった。

データベースアプローチは、このように従来は「例外」扱いされていた語や表現の関係を、表現レベルで網羅的に蓄積することで、例外を出さない言語間の対応（alignment）を実現している。

このアプローチで重要なことは、「節や名詞句の対応をどのようにして網羅的に収集するか」といういわゆる「辞書構築」の問題である。特に、上記の例でも示したように、ある言語で名詞句になっていても、翻訳文では、それを動詞句を使って表現している場合である。この問題を解決するために、名詞句に含意されている節（ここでは、「暗黙節」という。）を見つけ出すことである。これまでの研究で、隠れ節は、「形容詞＋名詞」、「名詞＋名詞」、「代名詞とその指示対象」、1つの動詞を共有する複数の名詞（主語や目的語）、「文の名詞化」、「接頭語＋名詞」、「名詞＋接尾語」などに節が隠されていることがわかっており、頻出するこのような名詞句に対しては、実際に名詞句から隠れ節を切り出した。この情報が、節対応(clause alignment)を自動的に見つけるのに重要な役割を果たす。

明示的な節や名詞句を自動的に抽出するために、日本語では Cabocha、英語では Stanford Parser を使い、これで解析した結果を更に、変換して節と名詞句を作成した。特徴的な点は、明示節（黒線）だけではなく、暗黙節（灰線）も切り出している点である。下2図は、1000 万件の英文から抽出した節と名詞句の数である。これから、総名詞句数、節数が推測できる。





これで総名詞句数、節数がわかる。1000 万文格納してもまだ少しずつ増加していく。