自然言語処理　84－16
(1991. 7. 19)

機械翻訳システムＤＵＥＴ　Ｅ／Ｊ　Ⅱの現状と今後の課題

鈴　木　等　　福　持　陽　士

シャープ株式会社・情報システム研究所

英日機械翻訳システムＤＵＥＴ　Ｅ／Ｊ　Ⅱは，ボトムアップ・シリアル方式の高速パーサ，２種類の意味体系を用いた強力な意味処理，タフな解析能力を持つ独自の２パス解析方式，約７，９０００語のエントリーを持つ基本語辞書などにより，高速，かつ，高精度の翻訳を実現している。本稿ではＤＵＥＴ　Ｅ／Ｊ　Ⅱの技術的特徴，及び，方式を概説し，現在の利用状況を紹介すると共に，現状の翻訳性能の限界を論じ，機械翻訳システムの今後の課題を展望する。

DUET E/J Ⅱ machine translation system: current situation and future hurdles

Hitoshi Suzuki　Yoji Fukumochi

Information Systems lab., Sharp Corp.

The English-to-Japanese machine translation system DUET E/J Ⅱ achieves high-speed and high quality translation by means of a range of technical features which include a bottom-up, serial high-speed parser; powerful semantic processing using two hierachies of semantic category; an original two-step parsing algorithm giving robust parsing; and a basic dictionary with approximately 79,000 entries. In this paper we shall outline the technical features and techniques used in DUET E/J Ⅱ and the way the system is currently being used. We shall also look at the present limits of its translation abilities, and the problems which machine translation systems have yet to overcome.

1．はじめに

「機械翻訳が使いものになりそうだ」という認識が機械翻訳の導入を検討しているユーザの間に広まってきている。このような認識が生まれてきたのは，実際に機械翻訳を利用して効果を上げている事例が多く現れてきたことによるものであろう。

昨今の国際情勢の目まぐるしい変化や，グローバリゼーションの進展に伴い翻訳の需要は確実に増大しており，機械翻訳に対する期待が高まっている。

本稿では，英日翻訳システムＤＵＥＴ　Ｅ／Ｊ　Ⅱの技術的特徴，及び，方式を概説し，現在の利用状況を紹介するとともに現状の翻訳性能の限界を論じ，機械翻訳システムの今後の課題を展望する。

2．ＤＵＥＴ　Ｅ／Ｊ　Ⅱの概要

2．1　技術的特徴

ＤＵＥＴ　Ｅ／Ｊ　Ⅱは，トランスファ方式を用いており，処理の大まかなフローは第１図に示す通りである。詳細については後節で説明し，ここでは技術的特徴を述べる。

（1）高速パーサ

構文解析部にはシリアル方式のボトムアップパーサを使用している。このパーサは，ルールをプリコンパイルして得られるＬＩＮＫテーブルの他，解析中に失敗したルールを記憶する機能や先読み機能などにより種々の高速化が図られている。

（2）構文解析と意味解析の融合

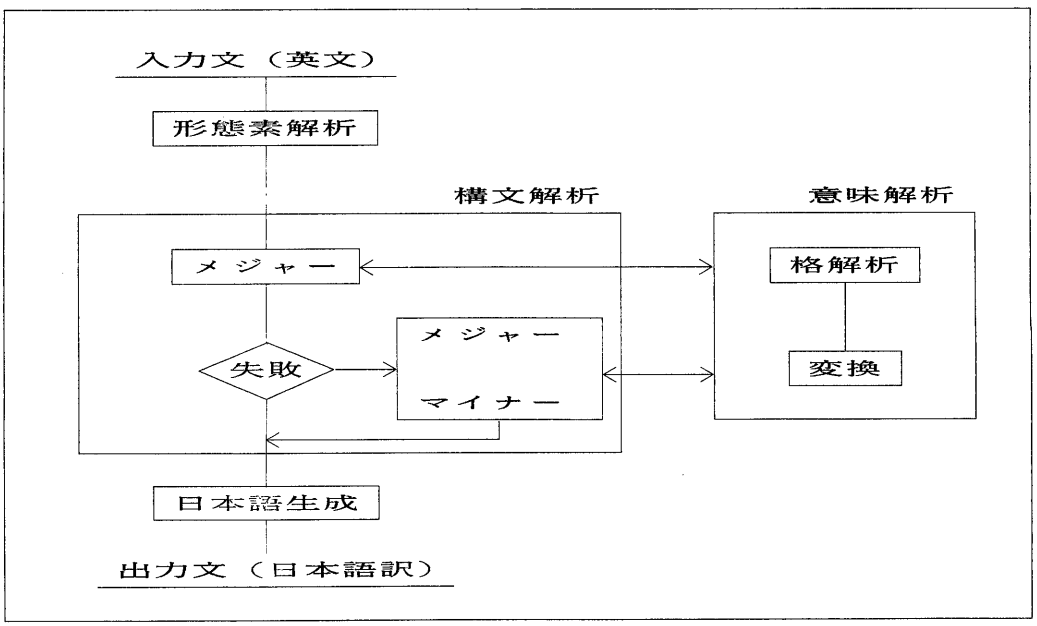
ＤＵＥＴ　Ｅ／Ｊ　Ⅱでは，構文解析中の任意の箇所で部分木単位の意味解析を行うことができる。このため，意味解析で失敗した場合，早い時点でバックトラックがかかるので，むだな解析を少なくできる。また，シリアルパーザの簡潔で見通しの良い制御も合わせ持つことができる。

（3）２種類の意味体系

意味解析で用いる意味マーカには，２種類のシソーラス体系を用いている。一方は格構造を決めるための体系であり，もう一方は訳語をうまく選択するための体系である。２種類の体系を用いているため，より柔軟な辞書記述ができるようになった。

（4）２パス解析方式

辞書は，各記述項目にプライオリティが付けられている。解析ルールについても同様である。これらのプライオリティのうちメジャーとマイナーの２値で規定



第１図　処理の概略フロー

されるプライオリティは，２パスの解析方式によって利用される。すなわち，１パス目はメジャーなものだけを使って解析し，失敗した場合はマイナーなものも利用する。この２パス方式によって，第１候補の翻訳結果の信頼性を高めるとともに，マイナーな表現にも対応できるタフな解析システムを実現している。

２．２　機械翻訳用辞書

ＤＵＥＴ　Ｅ／Ｊ　Ⅱは，見出し総数約79,000語からなる「基本語辞書」，ユーザによって構築される「ユーザ登録辞書」，専門分野別に構築された「専門用語辞書」を使用している。ここでは，上記のうち「基本語辞書」について述べる。

本辞書は(１)見出し情報，(２)文法情報，(３)訳語情報で構成されている。

（１）見出し情報
- （ａ）見出し番号
- （ｂ）代表見出し
- （ｃ）異綴り見出し
- （ｄ）確定フラグ／マイナーフラグ

複合語見出しが，いかなる英文中で用いられても，それ以上に分解した解釈があり得ない場合には，確定フラグを与えている。これによって，無理に分解して誤った解析が行われないようにしている。また，英単語は，複数個の品詞を持つことが多いが，使用頻度の低い品詞にはマイナーフラグを与え，１パス目の構文解析では排除するようにしている。

（２）文法情報
- （ａ）品詞番号
  多品詞語の優先度を与える。
- （ｂ）品詞名
  ３１種類に大分類されている。
- （ｃ）形態素情報
  語尾変化形，及び，形態的属性（人称，数，時制等）。
- （ｄ）品詞サブカテゴリ
  各品詞個別の細分類である。
- （ｅ）解析型
  各単語の持つ構文型で，品詞ごとにその必須格はパターン化されている。

（３）訳語情報

各単語に対して通常複数個の訳語が付けられており，各訳語は次のような情報が付加されている。
- （ａ）訳語番号
- （ｂ）訳語，及び，訳語の品詞
  当該語に対する日本語訳，及び，その訳語の品詞分解と活用型
- （ｃ）意味属性
  各訳語固有の意味属性であり，約30種類に大分類した「意味カテゴリー」と，約1000種類に細かく分類した「意味コード」の２種類の体系を用いている。
- （ｄ）訳語文法情報
  - （ｉ）用法
    当該訳語を選択するための文法的条件，日本語生成のための条件等を記述したものであり，構文解析，格解析，及び，日本語生成時に利用される。
  - (ii) 格パターン情報
    当該訳語が取り得る構文型の各々に対して記述される必須格条件であり，
    - （イ）必須格の表層格名（主語，名詞句，前置詞句等）
    - （ロ）深層格名（動作主，相手，道具等）
    - （ハ）生成型（日本語生成時に各必須格に付加される後置語，語順等をパターン化したもの）
    - （ニ）その他細部にわたる生成条件等で構成されており，格解析，及び，日本語生成時に利用される。
- （ｅ）格パターン意味情報
  個々の解析型における，必須格に対する意味上の制約条件が意味カテゴリー，または，意味コードにより記述されている。

２．３　形態素解析

形態素解析では，語尾処理により入力単語を原型に戻した後，辞書検索を行い，辞書に登録してある情報を引き出して構文解析に渡す。単語が辞書に登録されていない場合には，適当な処理を行って必要な情報を与え，翻訳処理を続行する。

２．３．１　語尾処理

入力原文の各単語について，三人称単数現在の“s”，過去形の“ed”，などを取って原形に復元する。
このとき原形の可能性のあるものをすべて候補としてあげる。

2．3．2　辞書検索

語尾処理で求めたすべての候補について辞書検索を行う。検索は，ユーザ辞書，専門用語辞書，基本語辞書の順に行われる。複合語は単語よりも優先して先に辞書から引いてくる。辞書に確定フラグがあれば，複合語を構成する個々の単語の辞書検索は行わない。

辞書検索を高速に行うために各辞書にインデックスを設けているが，特に基本語辞書では２段階のインデックスを設けている。

2．3．3　特殊処理

（1）ハイフン合成語（decision-support等）

ハイフンの前後の単語に対してそれぞれ適切な形態素処理を行う。これによって例えば

・a stem-mounted device

→　ステムを搭載したデバイス

・automatically-extended stack region

→　自動的に－拡張されたスタック領域

（2）その他

全体を一語として辞書に登録していなくても特定のパターンを持つ単語列は複合語とみなし，次のような情報を与える。

（a）数詞　two hundred and eighty-six

訳：２８６

（b）日付　July 1, 1980

品詞：名詞／副詞

訳：1980年7月1日

（c）時間　five mimutes past eight

品詞：名詞／副詞

訳：８時５分

2．3．4　未知語処理

辞書検索の結果，単語が辞書にない場合には，次のような情報を与える。

（1）接頭辞処理

接頭辞表に含まれる prefix（anti, re等）で始まる単語は接頭辞を除いた部分の辞書引きを行い，後で接頭辞の訳語と合成する。用言の接頭辞については単に接頭辞の訳語を付加するだけでなく次のような処理も行う。

・unexciting

不－興味をそそる　→　興味をそそらない

（2）接尾辞処理

接尾辞表に含まれる suffix（ness, ly, able等）で終わる単語は，表に記載された品詞（多品詞の場合もある）と訳語が与えられる。

・interrupt-driven

→　割り込み－駆動の

（3）算用数字

・1991

→　品詞：数名詞／数形容詞

訳：１９９１年，他

（4）記号を含む単語

・$50

→　品詞：名詞

（5）先頭が大文字で始まる未知語

・Sharp

→　品詞：固有名詞

2．4　構文解析

ここでは，構文解析部の制御機構であるパーサ，パーサが利用する構文解析用文法規則，構文解析部と意味解析部のインタフェースを成す木構造変換部について述べる。

2．4．1　パーサ

本システムの構文解析部では，入力文に対する品詞列を受け取って，句構造を出力するシリアル方式のボトム・アップ・パーサを採用している。同時に，処理の高速化を図るため，トップ・ダウンの予測機構を組み込み，失敗した規則を記憶することで，無駄な解析をできるだけ行わないようにしている。また，解析の確度を高めるために，文法規則に記述されたマイナーな規則であることを示すフラグや辞書に記述されるマイナー品詞フラグ，及び，マイナー構文型を参照して，マイナーな解釈を一番目の解釈としては出力しないように，２パス方式（１パス目では，メジャーな解釈を用いて解析し，失敗した場合のみ，マイナーな解釈を出力する方法）を採用している。

また，構文解析と意味解析を完全に分離してシリアルで処理を行うと，意味解析によって明らかに排除できる条件が辞書に記述されていても，無駄な構文解析を行うことになるので，ＤＵＥＴ　Ｅ／Ｊ　Ⅱの構文解析部では，必要とするノード（例えば，動詞句や文）

にたいして，随時，意味解析を実行できるような機構を備えている。

パーサが実行する文法規則は，書き換え規則の各項に，補強項としてのチェック関数が記述できる拡張文脈自由文法で記述されており，規則の増大を押えるために，ＶＰ／ＮＰのようなスラッシュ・カテゴリーが使用できるようになっている。その詳細は次の通りである。

２．４．２　構文解析文法

ＤＵＥＴ　Ｅ／Ｊ　Ⅱには，現在，約1,200個の文法規則が存在し，補強項として書かれたチェック関数では，大きく分けると次の４つのチェックを行っている。

（１）統語的に非文となるような解釈の排除

例：辞書中に記述された解析パターン，品詞細分類，形態属性（名詞句や述語の数など），解析部で与えられた統語属性をチェックする。

（２）ノードへの特定の属性，及び，属性値の付与

例：Ｖ　→　ＢＥ　＋　ＶＥＲＢ（過去分詞）の書き換え規則において，Ｖのノードに様相＝受身の情報を与える。

（３）品詞や係り受けについて曖昧性が生じた場合，ヒューリスティックな知識を用いた解釈の優先

例：～ set the values and changes parameters

下線の部分の解釈として，並列名詞句の解釈は，排除できないが，名詞（複数形）＋名詞の名詞連続よりも“change”は，動詞としての解釈を優先する。

（４）規則適用時に後方に要求される品詞などの先読み

例：～ asked the man to do the job.

ＤＵＥＴ－Ｅ／Ｊ－Ⅱでは，ホンビーの動詞型[4]を拡張したものを使用しており，上記の入力文では，

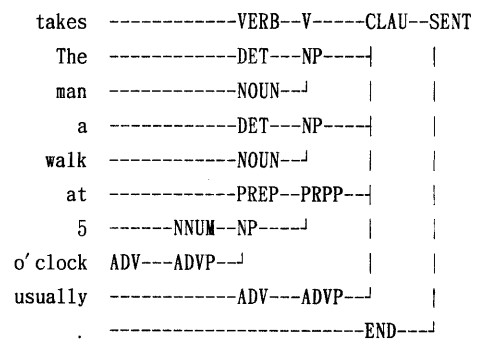
ＶＰ　→　Ｖ　ＮＰ　ＩＮＦ

という文型パターンの構造をしているが，Ｖのノードから，後方に不定詞となり得る品詞列があるかどうか先読みすることで，無駄な解析を行わないようにしている。

２．４．３　木構造変換部

一般的に，自然言語を句構造に解析すると，意味的には同じでも様々な構文構造木を出力してしまう。そのため，この木構造変換部では，この構文木を正規化し，後に続く意味解析が行いやすい構造に tree-to-tree の変換を行う。したがって，木変換規則は，文脈依存文法で記述できるようになっており，規則，及び，パターンマッチの制御も行えるようになっている。入力文“The man usually takes a walk at 5 o'clock.”に対して木構造変換を施した内部構造を第２図に示す。

The man usually takes a walk at 5 o'clock.



第２図　ＤＵＥＴ－Ｅ／Ｊ－Ⅱの内部表現

２．５　格解析

ＤＵＥＴ　Ｅ／Ｊ　Ⅱの格解析部では，構文解析で得られた構文構造を基本語辞書に記述された格パターン意味情報と照合してふるいにかけると同時に，成功した構造に対して構文，及び，訳語の変換も行ってしまう。

第１ステップでは，

（１）文型パターンのチェック

（２）格要素のチェック

（ａ）各種用法のチェック

（ｂ）表層構造のチェック

（ｃ）意味カテゴリー制限を満たしているか。

（ｄ）共起語を要求している場合，その語があるか。

を行うことによって，訳語の候補を絞り，更に，第２ステップで

（３）深層格のチェック

（４）詳細意味コードのチェック

を行う。

表層構造は，単語や句の間の文法的な係り受け関係であり，深層構造は，意味的な役割関係を表す。

任意格の深層格は一意に決まらないことも多いが，既に必須格として同一の深層格を取っている場合には他の深層格を与えることにより「～は，～に～に～する」といった同一格の重複を排除している。

格解析に失敗した場合は，再び構文解析にバックトラックし，他の構文構造がないかどうかが調べられる。

2．6　日本語生成

格解析が終了すると，各単語の訳語，中心語に対する深層格，後置語（格助詞，及び，格助詞相当語），様相などが確定している。

日本語生成処理部では，生成用テーブルを参照しながら，これらをある一定の順序で並べることにより日本語を生成していく。生成用テーブルには，深層格と，後置語の対応等が入っており，深層格によって訳語の順序が決定し，同時に，訳語に付加する格助詞等も決定する。

（1）必須格処理

必須格は，生成用テーブルを参照しながら生成していく。ただし，辞書に生成順序の指定がある場合には，辞書の順序が優先される。

まず，生成用テーブルを順番にたどりながら，テーブルの深層格と格解析で得られた深層格とが一致するかどうかを検索していく。一致すれば，その深層格をもつ中心語と，その中心語に係る語を生成する。このとき，その中心語の品詞に応じた処理を行うことにより，自然な日本語を生成している。また，この中心語が様相格をもっていれば，様相の処理を行う。

次に，その訳語に対して深層格に応じた後置語を付加する処理を行う。生成用テーブルでは予め深層格に応じた後置語が定められているが，より自然な日本語を生成するために辞書の中に後置語を指定している場合は辞書に記述されている後置語を用いる。

（2）任意格処理

訳文における任意格の生成位置は，原則として，原文（英文）に現れる位置をもとに決めている。すなわち，文頭の副詞句等は，訳文でも文頭に置き，その他の副詞句，前置詞句等は被修飾述語との距離が原文に近くなるように定められる。

3．DUET　E／J　Ⅱの利用状況

ここでは昨年１２月に当社のDUET　E／J及びDUET　E／J　Ⅱのユーザを対象に行ったアンケート調査の一部を紹介する。

（1）翻訳文献の種類（複数回答）

a)マニュアル	19.7%
b)技術文献	18.8%
c)雑誌	12.0%
d)海外レポート	11.1%
e)海外レター	8.8%
f)カタログ	7.4%
g)特許関連文書	6.0%
h)英字新聞	3.7%
i)テレックス	3.1%
j)海外規格	3.1%
k)不動産情報	1.1%
l)その他	5.1%
有効回答	351件

翻訳文献の種類は昨年６月に行った調査とほぼ同じ傾向を示しており現状の機械翻訳の利用分野を反映していると思われる。

（2）主な使い道

a)一括翻訳の後，手を加えて訳文を直している（下訳）	60.8%
b)英語文献の大意を取るための一括翻訳（概要把握）	30.8%
d)その他	8.5%
有効回答	130件

a)下訳と b)概要把握の割合は約２：１となっており，以前に比べて後者の割合が増えてきている。電子化文書の増加，翻訳システムの低価格化により，この傾向は今後益々増えて行くものと思われる。

（3）使用者

a)部署の全員が必要に応じて使用	31.5%
b)専任のオペレータ	25.0%
c)会社の全員が必要に応じて使用	15.3%
d)OCR入力，翻訳を別々の人が分担	8.1%
e)その他	20.2%
有効回答	124件

a)と c)の不特定多数による使用の合計はは46.8%と圧倒的に多く，パーソナル的な利用が進んできていることを示している。一方，稼働率が高いのは，b)の専任オペレータや d)の分業体制による利用である。

（4）利用形態

a)単独で利用	71.8%
b)パソコンとの連携	17.9%
c)その他	10.5%
有効回答	117件

約30%が何らかの形で他の機器と連携して使用している。これは，既存のＯＡシステムの中に組み込むことによって従来の人手を中心とした翻訳作業の効率化を目指しているものと思われる。

（5）機械翻訳利用の効果

現状ユーザが機械翻訳に対して感じている又は期待しているメリットとしては次のような項目がある。

高速翻訳	⇨ ・翻訳作業のスピードアップ ・処理量の増大
翻訳作業の分業化	⇨ ・翻訳作業の効率アップ ・労働力の確保
翻訳品質の統一（特に専門用語の訳し方）	
機密の保持	
ユーザ辞書の構築による知識の資産化	
メディアによる入出力	⇨ ・データベース化
大量海外情報の取捨選択	⇨ ・英文速読支援 ・死蔵海外情報の活用

利用事例

（a）翻訳工場

・OCR入力，前処理，後処理，リライト，を分業化
・コンピュータマニュアル専門
・ＤＵＥＴとパソコンの連携
・１人当たりの生産量が120枚/月（Ａ４版の英文で換算）から機械翻訳利用により180枚/月に増加
・専門用語辞書による訳語の統一
・労働力の確保

（b）外電速報

・英日対訳形式でキーワードのみを和訳
・時事専用のキーワード辞書を作成
・毎日一定時に所内に回付

（c）機械翻訳出版[5)]

・原文を前処理し，機械翻訳したものを下訳として利用
・機械の出力した下訳をもとに翻訳者が最終的な日本語へ翻訳

（d）翻訳者の養成

・機械翻訳機を用いた翻訳者養成スクール
・翻訳機のオペレーションだけを教えるのではなく翻訳のテクニックを指導

5．機械翻訳の限界

機械翻訳の２大利用法である「下訳」と「大意把握」で真に効果を挙げているユーザが増えてきており，このことが機械翻訳は「翻訳業務の効率向上を支援するツール」であるとの位置付けを明確にしてきている。

しかし現在の機械翻訳にユーザは必ずしも満足しているわけではない。より高精度で，より安価で，より使いやすいシステムを望んでいる。

導入検討ユーザが挙げる問題点には，

1）翻訳品質
2）システム価格
3）操作性
4）ファイルの互換性
5）後編集
6）ユーザ辞書作成労力
7）前編集
8）処理速度
9）要員の確保
10）辞書品質
11）入力手段
12）システム拡張性
13）サポート体制

などがあり，1)の翻訳精度が常にトップに位置している。翻訳精度に対するユーザの不満は永遠のものであると思われるが，現在の機械翻訳には，

1)短文はまずまずだが長文がうまく訳せない
2)１文単位の翻訳のため文章としての一貫性がない
3)雑誌の見出しのような凝った表現は意味不明の訳になることが多い
4)機械翻訳に適した文章の領域が狭い

などの弱点があり，ユーザはこれらの弱点を知った上でツールとしての翻訳システムに期待している。

6．今後の課題

現在の一文単位の翻訳に対して，文章の前後の意味的なつながりをくみ取る，いわゆる文脈処理技術を応用した機械翻訳が考えられているが，広い範囲をカバーする実用的な機械翻訳に文章理解技術を適用するには，まだ時間がかかる。現在の技術の延長として我々がＤＵＥＴ　Ｅ／Ｊ　Ⅱで取り組もうとしている課題としては，

- 長文の解析処理の高速化
- プライオリティや経験的知識を用いて多数の解釈の中から，より妥当性のある解釈を選択すること
- 翻訳者や、テクニカルライターの翻訳ノウハウ，文章表現ノウハウを取り込んだ自然な訳文生成
- 分野ごとの文章的特徴を取り込んだ精度の向上
- 大量の慣用表現の取り込み
- テキストのレイアウト情報を用いた文種の認識

などが挙げられる。また，近い将来に向けての取り組みとして

- 特徴語や頻出単語等を用いた分野／テーマの自動特定[6)]
- 省略の補充，挿入文や倒置の認識

などがある。これら言語処理の改善による翻訳精度の向上のほかに，

- ＯＣＲ読み取り後の後処理の自動化
- 前編集テクニックの体系化と自動化
- readabilityの改善を目指した翻訳結果の後処理の自動化

等，作業効率を上げるための周辺ソフトの充実が必要である。また，

- 専門用語データベースの構築とその流通
- 機械翻訳向き制限言語の標準化とその普及
- 翻訳システムの操作やファイルの規格統一
- 機械翻訳システムの評価基準の設定

といった機械翻訳を取り巻く環境の整備についても考えていく必要がある。この４月に設立された日本機械翻訳協会（長尾真会長）は，開発者だけでなく，利用者，サービス業者，研究者が協力して機械翻訳の健全な普及，発展を図ることを目指したもので，今後の活動が期待される。

7．むすび

本論文では当社の英日機械翻訳システムＤＵＥＴ　Ｅ／Ｊ　Ⅱの技術的特長と仕組みを概説した。また，ユーザアンケートの一部を紹介することにより機械翻訳システムの利用状況を述べた。さらに，現状の翻訳性能の限界を論じ，将来の機械翻訳へ向けての課題を展望した。機械翻訳の歴史は古く1930年代にさかのぼるが，初期の楽観的時代からＡＬＰＡＣレポートの見直しの時代を経て，今ようやく実用化の時代を迎えた。今後更に翻訳精度向上の努力と共に，低価格化によるパーソナル分野への普及が期待される。

参考文献

1) 野村浩郷・田中穂積編：「機械翻訳」，共立出版　ｂｉｔ別冊（1988年9月）
2) 山田博編：「人工知能ハンドブック」，オーム社
3) 鈴木等他：「英日機械翻訳システム DUET-E/J」，シャープ技報第41号（1989年）
4) ホーンビー：「英語の型と語法」，オックスフォード大学出版局（1977年）
5) 小宮山信之編：「対訳 ブッシュ大統領決断のスピーチ」，中経出版（1991）
6) Stephen I. Gallant: "A Practical Approach for Representing Context And for Performing Word Sense Disambiguation Using Neural Networks", NU-CSS-90-5 (1990)