デジタル・ドキュメント 16－2
（1999. 1. 29）

情報化社会における多言語解析とインターネット
－日中、日韓における翻訳メールシステム－

河野 勝也　松田 純一　隈井 裕之
(株)日立製作所 中央研究所

インターネットに代表される情報化社会の世界的な進展を踏まえ、我々は個人間の母国語によるコミュニケーション支援を目標に多言語情報処理の研究に取り組んできた。本報告では、筆者らが開発した中国語文章入力システムと、日中、日韓翻訳メールシステムについて報告する。中国語文章入力システムは、中国語の英字表記であるピン音表記の入力を文章単位で中国語に変換する。変換の過程で生じる同音語の問題を、中国語の特性に着目した文法ルールを開発することで解消し、92%の変換率を実現した。日中、日韓翻訳メールシステムは、原文と翻訳結果を同一画面に併記してメールを送信できるようにし、受信者は誤訳等で意味が不明確な場合でも、原文にあたることで内容を理解できるようにした。

Japanese-Chinese and Korean Machine Translation E-mail System

Katsuya KOHNO, Junichi MATSUDA, Hiroyuki KUMAI,
Hitachi,Ltd., Central Research Laboratory

We have been pursuing the research on multi-language processing technology to enable people of different languages to communicate with each other in their native language, in an increasingly global information society. A Chinese word input system, and Japanese-Chinese (J-C) and Japanese-Korean (J-K) machine translation electronic mail system are introduced in this report. The Chinese word input system we developed, changes pinyin romanized Chinese into Chinese characters, sentence by sentence. To discriminate pinyin as different Chinese characters, which is the most difficult problem, we developed special grammatical rules based on Chinese language characteristics which successfully allows 92% of the pinyin to be transformed into the correct Chinese characters. The J-C and J-K translation e-mail system allows the original sentence and the translated sentence to be displayed on the same screen so that in the case of a mistranslation, the recipient can refer to the original text.

1.はじめに

先進国を中心に進んできた情報の電子化の流れは、パーソナルコンピュータの本格的登場以降、世界中で推し進められるようになってきた。一方、情報ネットワークは世界的な広がりを持つに至り、特に電子メールは、国境を越えた個人のコミュニケーションを強力に支援する手段となっている。

このような情報化社会の進展の状況を踏まえ、我々は個人間の母国語によるコミュニケーションの支援を目標に多言語情報処理の研究に取り組んできた。

本報告では、多言語でのコミュニケーションの円滑化を図るという目標のもと、筆者らが取り組んできた研究の一部である、中国語文章入力システムと、日中、日韓翻訳メールシステムについて詳しく述べる。

多言語情報処理の課題として、多言語文字の入力、表示、情報交換が挙げられる。

図 1.1 に示すように、英語圏を中心に発達してきたコンピュータは、早くからアルファベットの入力と表示をサポートしてきた。一方、日本語については、コンピュータの処理性能の向上によって、ビットマップディスプレイや、大容量の文字フォント、カナ漢字変換を始めとするフロントエンドプロセッサがの登場によっ

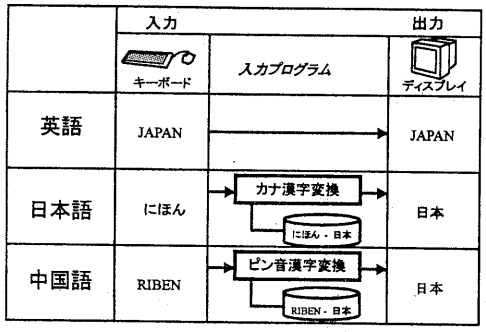

図 1.1 多言語文字処理

て漢字を含む日本語の入力、表示が可能になった。また、中国語に関しても後述するように、筆者らが開発したピン音漢字変換を用いた複数文節中国語文章入力システムシステム等を用いることにより専門オペレータでなくとも、容易に中国語の入力が可能とすることができるようになった。

情報交換に関しては、当初アルファベットのみであったインターネットメールでも、日本語が扱えるようになり、日本人どうしは日本語で電子メールのやり取りができるようになった。

同様に、非アルファベット圏の中国や韓国でも、電子メールが普及してきており、各国内では中国語どうし、韓国語どうしの電子メールのやり取りが行われている。今後、東アジアの文化、経済の交流が進む中で、日本と中国や韓国との間で電子メールのやり取りも増加すると思われる。

現在、メール交換に使用されるパソコンは、

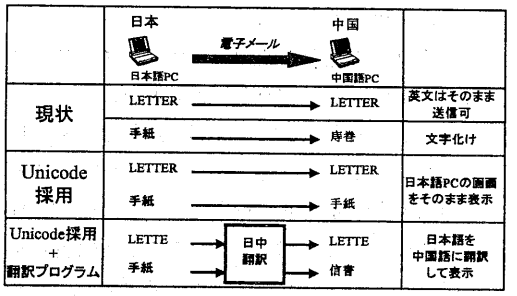

図 1.2 電子メールにおける文字コードの問題

Unicode	日本 SJIS Code	中国 GB Code	台湾 Big5 Code	韓国 KS Code
9AA8	骨 8D9C	骨 B9C7	骨 B0A9	骨 CDE9
5203	刃 906E	刃 C8D0	刃 A462	刃 ECD3
673A	机 8AF7	机 BBFA	机 C9F3	机 CFF5
5F10	弐 93F3			

図 1.3 Unicode の例

非英語圏では自国語と英語の２ヶ国語対応になっている。このため、英語以外の言語での国際間のメール交換は、事実上不可能になっており、図 1.2 の現状に示すように、日本語のメールを中国で読むことさえ困難であり、日中韓は同じアジア漢字圏であるにもかかわらず、電子メールの交換は英語に頼らざるを得ない状況にある[1][2]。

この問題は、各国の文字コード、独自に規格化されてきた経緯に端を発するものである。

近年、この問題を解消すべく、Unicode をベースとした国際符号化文字集合 ISO/IEC 10646-1 が規格化された。図 1.3 に Unicode の一部を例示する。Unicode には、同じコードでも字形が各国によって微妙に異なるなど問題がないわけではないが、Unicode を採用する汎用のオペレーティングシステムも登場し、多言語情報処理の基盤が整いつつある[3]。

我々は、このような状況に鑑み、日本と中国・韓国間で母国語を用いて相互にコミュニケーションできることを目指し、日中・日韓双方向間の機械翻訳メールシステムを開発した。

本報告では、日中間の相互コミュニケーションを中心に、複数文節中国語文章入力システムと日中、日韓翻訳メールシステムの開発について報告する。

2 複数文節中国語文章入力システム

2.1 開発の経緯

従来、中国語入力方式では、四角号碼法、五筆法など漢字を分解してコード化して入力する方法が多く提案されている[4]。

このような入力方法は、相当の学習を必要とするため、専門オペレータが原稿を見ながら入力するのには適しているが、文章を考えながら入力する場合には、思考の妨げとなる。

そこで、ユーザの思考を妨げない入力には読みからの入力が最適と考え、中国語の英字表記であるピン音を用いた複数文節中国語文章入力システムを開発した。本システムはパーソナルコンピュータ上のフロントエンドプロセッサとして稼動する。

2.2 システムの概要

ピン音は発音表記に近く、従来のコードや字画、部首を用いた入力方法に比べ、初心者でも容易に効率的に入力できる。

本システムは、図 2.1 に示すように中国語を文単位に読みで入力し、中国語漢字文字列に変換するものである。中国語用語辞書は約 5 万 5 千語の用語を持ち、文法解析に必要な品詞情報を持つ[5]-[9]。

読みを入力して漢字変換するときの技術課題は日本語のかな漢字変換と同様に同音語の中から、いかに適切な漢字を選択するかにある。

この問題を解決するためには、①文法知識の応用による漢字の特定化、②漢字の使用頻度情報の利用、が有効である。

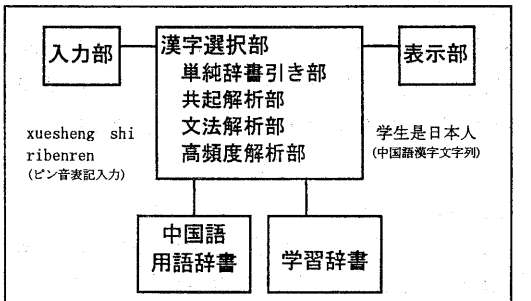

図 2.1 中国語文章入力システムの概要

中国語：学生买一本书。（学生は一冊の本を買う。）					
ピン音入力:	XUESHENG	MAI	YI	BEN	SHU.
①単純辞書引き:	学生	MAI	YI	BEN	SHU。
②共起解析:	学生	MAI	YI	本	书。
③文法解析:	学生	MAI	一	本	书。
④高頻度解析:	学生	买	一	本	书。

図 2.2 変換経過の例

2.3 漢字選択の方法

以下、同音語を高精度に選択する漢字選択部の動きについて説明する。

キーボードから中国語の読みが標準ピン音表記で文単位に入力されると、以下の手順で処理する。

①単純辞書引き部：単純に辞書検索を行って、読みに同音語がなければ漢字を決定。

②共起解析部：共に現れる可能性の高い読みが現れれば漢字を決定。

③文法解析部：単語の位置および両隣の単語の品詞の並びの特性を、文法ルールとして定義し、文法ルールに一致したとき、品詞または漢字を決定。

④高頻度解析部：上記で決定できない場合、同音語中最も使用頻度の高い漢字で決定。

変換経過の例を図 2.2 に示す。

①XUESHENG には同音語はなく辞書引きだけで「学生」に決定する。

②BEN には 9 個、SHU には 46 個、同音語があるが BEN SHU が同時に現れたときには、共起解析部において量詞と名詞の組合せである「本」「書」の可能性が高いと判定し漢字を決定する。

③YI には 110 個同音語があるが、「本」が量詞で決定されているため、量詞の前には数詞が来るという文法知識（ルール）を用いて、「一」

に決定する。
④上記で決定できなかったMAIには9個の同音語があるが、辞書に格納されている頻度情報を参照し、最も使用頻度の高い漢字で決定する。

2.4 文法ルール

文法解析で用いられる文法ルールには、中国語の文法知識に基づいて、以下のようなルールを設けている。

(1)文頭・文末解析ルール

文頭、文末の位置情報を用いて解析する。ルールの例を図2.3のR1からR3に示す。

(2)漢字決定両どなり解析ルール

前後に隣接する単語の品詞や特定の読みなどに注目し、漢字を決定する。一例として、助動詞「DE」の同音漢字「的」「得」「地」の選択し分けのルールを図2.3のR4からR7に示す。

(3)品詞決定両どなり解析ルール

前後に隣接する単語の品詞や特定の読み、あるいは既に決定した漢字などに注目し、品詞を決定する。もし、決定した品詞に漢字が複数ある場合には、最終的に高頻度解析で漢字を決める。図2.3のR8からR12に例示する。

R1:文頭で「，」が続くとき、文頭は感嘆詞あるいは助詞とする。
R2:文頭で「，」が続かないとき、文頭の品詞に代詞があれば代詞とする。代詞がなく介詞があれば介詞とする。
R3:文末の読みの品詞に助詞があれば助詞とする。

R4:読みがDEで前が名詞または代詞のとき「的」で決定
R5:読みがDEで後が名詞または「。」のとき「的」で決定
R6:読みがDEで後が動詞で前が形容詞のとき「地」で決定
R7:読みがDEで後が動詞で前が形容詞でないとき「得」で決定

R8:前が量詞なら後は名詞
R9:後が量詞なら前は数詞
R10:後が動詞なら前は副詞、形容詞
R11:前が助動詞なら後は動詞、形容詞
R12:前が介詞なら後は名詞 代詞 数詞

R13:同じ読みが重なっているとき 動詞か 形容詞か 名詞
R14:「一」または「了」をはさんで同じ読みが重なっているとき 動詞か 形容詞

図2.3 文法ルール

(4)読み特性による品詞決定両どなり解析ルール

中国語では、同じ漢字を繰り返すことによって意味を添えるものがあり、よく使用される。前後の漢字が決定されていれば、繰り返される漢字の品詞を決定することができる。図2.3のR13からR14に例示する。

2.5 べた書きピン音入力への対応

ピン音の正書法では、単語単位で区切りを入力する。具体的には、単語と単語の間にスペースを挿入する。本システムでは、複数文節の入力の際に、ピン音の間に入力されたスペースは、単語の区切りとして扱うこととした。

しかし、日本語のカナ漢字変換入力から連想されるように、複数文節の中国語のピン音をべた書きで入力することを望むユーザも多い。そこで、本システムは、べた書きで入力されたピン音も複数文節の中国語に変換できるようにしている。

べた書きで入力されたピン音を中国語に変換する場合、単語の複数の区切り位置の可能性、すなわち単語区切りの多義が生じ、ユーザがスペースで区切った場合に比べ、変換率は低下する。正しい変換結果を得るためには、正しい区切り位置を選択する必要がある。

ピン音を連続して入力された場合の、単語区切りの多義の例を図2.4に示す。

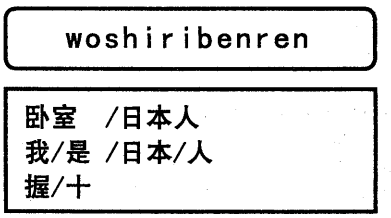

図2.4 単語区切りの多義の例

日本語のカナ漢字変換では、最長一致法、文節数最小法等のアルゴリズムによって、べた書き入力されたカナ文字列を概ね正確に複数の文節に区切ることができる。これは、文節はより長いほうが日本語として尤らしいというヒューリスティックルールに基づく。

しかし、中国語では、ピン音長の長くなる単語の分割よりも、ピン音長の短くなる単語の分割の方が正しいケースが多々ある。この理由は、中国語では、特に基本単語には1語の単語が多いためである。

また、日本語が、付属語が存在し音読みと訓読みが繰り返されるなど文節の区切りが推定しやすいのに対し、中国語は、単語の分割位置を示すような特定のピン音はなく、また、冗長性が少ない。

単語区切りの多義を解消するには、先に述べた共起解析や文法解析を、単語区切りのすべての組み合わせに対して行うことも有効と考えられるが、すべてのルールを適用すると、変換時間が実用の範囲を超えることが予想された。

本システムでは、中国語のピン音表記の性質に鑑みて、べた書きピン音入力に対して、より高速に単語の区切り位置を決定できるように、次の方法を取り入れた。

中国語用語辞書の各単語を、常用的に使用される常用語と、それ以外の非常用語に分類し、更に、ピン音の先頭から単語区切りの位置を決定する以下の単語区切り処理を追加した(図2.5)。

①入力されたピン音文字列に対して、中国語用語辞書を検索し、すべての単語区切りの組み合わせを生成する。

②単語区切りの組み合わせに対し共起解析部のルールを適用し、単語区切りを決定。

③②で適用するルールがない場合、常用語と非常用語が存在する場合、常用語のピン音長を優先する。常用語のみ、非常用語のみが存在する場合には、ピン音長の長いものを優先し、単語区切り位置を決定する。

④2.3節で述べた「漢字選択の方法」にしたがって同音語を選択する。

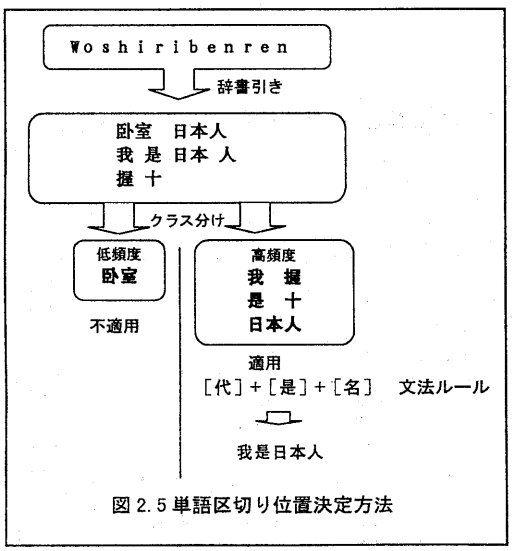

図2.5 単語区切り位置決定方法

2.6 評価

以上の解析処理を組込んだ中国語入力システムを開発した。表2.1に、べた書きのピン音入力とした場合の、各分野19文書(16826文節)を対象とした、変換率評価結果を示す。結果は、

表2.1 評価結果

分類	番号	文節数	正変換数	正変換率
文法書	1	2888	2262	78.3%
	2	5132	4185	81.5%
文学・思想	1	828	645	77.9%
	2	1264	908	71.8%
パソコン関連	1	678	540	79.6%
政治・経済	1	920	651	70.8%
	2	349	261	74.8%
	3	838	599	71.5%
理工学	1	355	254	71.5%
教育スポーツ	1	73	46	63.0%
	2	465	338	72.7%
	3	277	208	75.1%
	4	648	464	71.6%
その他	1	345	242	70.1%
	2	493	216	43.8%
	3	731	473	64.7%
	4	247	188	76.1%
	5	125	82	65.6%
	6	170	136	80.0%
合計		16826	12698	75.5%

全文節数に対する正しく変換された文節数の割合を正変換率として示している。全文書平均で75.5%の正変換率が得られた。2.5節で述べた単語区切り処理を行わなかった場合には、58%であった。

なお、単語単位に区切ってピン音を入力する場合の評価では、文法ルールを用いた場合には92%、用いなかった場合は87%であった。

以上の結果から、本システムの文法ルール、単語区切り方式の有効性が確認できた。本システムが、実用に供しうる変換率を実現できる見通しを得た。

3.翻訳メールシステム

日本と中国・韓国間で母国語を用いて相互にコミュニケーションできることを目指し、図3.1に示す日中・日韓双方向間の機械翻訳メールシステムを開発した。

3.1 システムの機能

図3.2は、中国にメールを送る場合の本システムの画面例である。

ユーザは原文作成エリア(①)に、日本語のメール文章を作成する。次に、翻訳先の言語(中国語、韓国語)を指定する(②)。ここでは、日本語→中国語、日本語→韓国語、中国語→日本語、韓国語→日本語が選択できる。次に、翻訳を実行(③)すると、翻訳結果表示エリア(④)に、原文と翻訳結果が文単位に交互に表示される。次にメール本文の生成を実行(⑤)すると、翻訳結果表示エリアの内容が、HTML形式のメール本文が生成され、メール送信画面が表示される(図3.3)。この時、生成されたメールには、原文

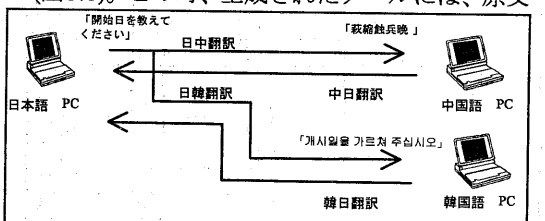

図 3.1 翻訳メールシステム概要

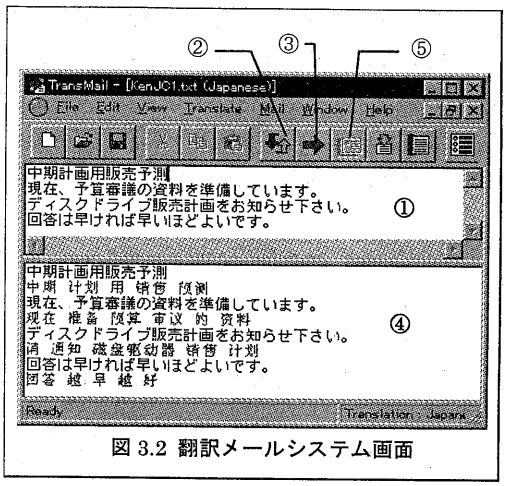

図 3.2 翻訳メールシステム画面

と翻訳結果が同一画面に併記される。ユーザは、宛先と件名を指定しメールを送信する。

受信者は、本システムの受信メール一覧表示、メール内容表示機能を用いて、図3.3と同様な原文と翻訳結果が併記されたメールを表示することができる。

原文が併記されているため、翻訳結果に誤訳があったり、意味が不明確な場合には、原文にあたることで、受信者はメールの内容を理解できると期待できる。

日本語から韓国語に翻訳したメール(図3.5)、中国語、または韓国語から、日本語に翻訳したメールも、同じ操作で送信できる。

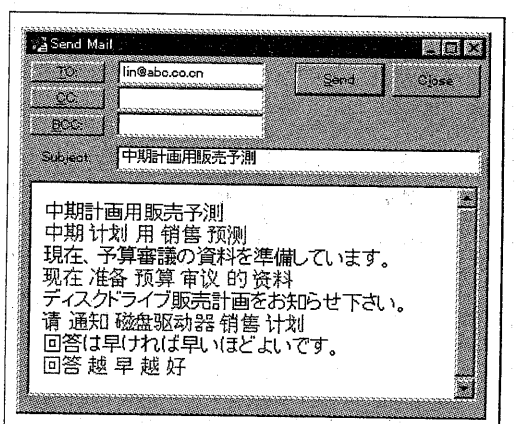

図 3.3 メール本文画面

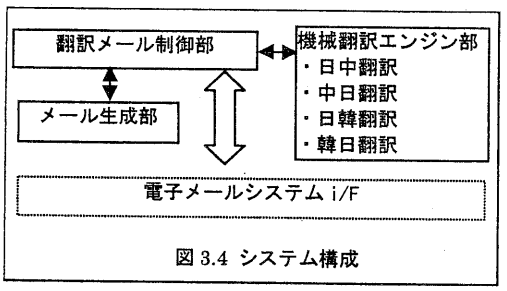

図 3.4 システム構成

3.2 システム構成

本システムの構成を図3.4に示す。

(1)翻訳メール制御部

翻訳メール制御部は、主にユーザi/Fを担う部分であり、翻訳対象の原文の入力、翻訳の実行、翻訳結果の電子メールへの変換、メールシステムの標準i/Fを通したメール送受信の機能を担う。

(2)機械翻訳エンジン部

日中翻訳エンジンは、意味トランスファ方式を基本に、日本語と中国語の依存構造が異なる場合でも的確な翻訳文を生成できるよう慣用句表現への対応処理を加えたものである。

日韓、韓日機械翻訳エンジンは日本語と韓国語の特性を考慮し、構文ダイレクト方式[10]を用いたものである(図3.5)。

(3)メール生成部

メール生成部は、原文と翻訳結果から、HTML形式のメール本文を合成、生成する。

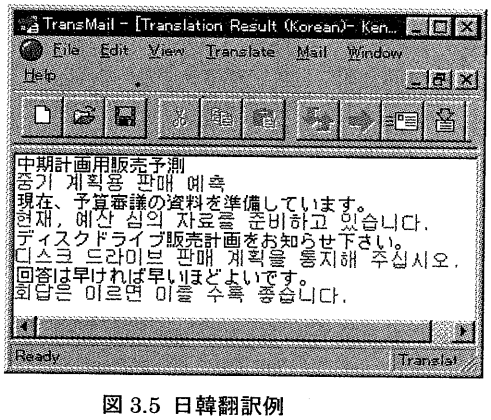

図 3.5 日韓翻訳例

4.おわりに

多言語情報処理研究の一環として、中国、韓国との間で母国語による相互コミュニケーションを実現する、複数文節中国語文章入力システムと日中、日中・日韓双方向間の機械翻訳メールシステムを開発した。

複数文節中国語文章入力システムについては、評価結果を解析し、ルールの適用条件の妥当性を検証し、変換精度を更に向上させることが今後の課題である。

翻訳メールシステムについては、日中、日韓間におけるフィールドテストを行い、本システムの評価と、メール文における機械翻訳の精度向上が、今後の課題である。

［参考文献］

[1]西垣：多言語時代を迎えたインターネット；世界 第640号(1997.10) 岩波書店

[2]三浦：多言語主義とは何か；藤原書店(1997)

[3](財)国際情報化センター国際規格共同開発調査:多言語情報処理環境技術成果報告書(1998)

[4]陳他：中国語の漢字入力の一方法；情処学会第35回全国大会

[5]香坂：現代中国語辞典；光生館

[6]三野：中国語文法の基礎；三修社

[7]朱、杉村他訳：文法講義；白帝社

[8]王他、林訳：中国語動詞活用辞典；東方書店

[9]高橋他：中国語虚詞類義語用例辞典；白帝社

[10]松田、河野：構文ダイレクト方式による日韓機械翻訳システム；情処学会第44回全国大会