機械翻訳システム用日本語OCR

3E−4

伊藤悦雄　武田公人
(株)東芝　総合研究所

1. はじめに

機械翻訳システム(MTS)の原文入力手段として、OCRが広く用いられるようになってきているが、両者を単に接続させただけでは、インターフェースが不統一であったりデータの転送時などにOSのコマンドを使用する必要があるなど、使いやすいシステムにはなり得ない。

そこで、我々は、MTSの原文入力機能という観点から、OCRにはどのような機能が要求されるかを検討した上で、MTSの原文入力に適した日本語ソフトOCRを開発した。

本稿では、今回開発した機械翻訳システムASTRANSACの原文入力用日本語ソフトOCR(AS-Reader J)の概要とその機能に付いて述べる。

2.機械翻訳システム用OCRへの要求

OCRを接続したMTSへの要求として「読み取った原稿のイメージ(レイアウト・フォントなど)をそのまま保持した形式の訳文が人手を介すこと無く得られる」がある。また、オペレータが介在する場合にも、システムに一貫性が要求される。

これらの要求は具体的には以下のようになる。

・レイアウト保存機能

原稿のレイアウトを保持し、機械翻訳システムへ渡せる機能が必要である。

OCRにこの機能があれば原文のレイアウト情報を訳文に反映させる翻訳システムは開発されているため[1]、原稿のイメージを持った訳文を得ることができる。

・「文」の認識

従来のOCRでは認識結果を「文」として処理する必要がなく「文字」として扱っていた。しかし、MTSでは、「文」を単位として翻訳を行うため、「文」を認識する必要がある。

・操作の一貫性

MTSとOCRとの操作性の統一が必要である。また、原稿読み込みから翻訳・印刷までを最少の操作で行うために、OCRの操作画面から機械翻訳への文書登録や翻訳開始指示、翻訳パラメータの設定などが行うことが要求される。

・編集機能の統合

OCRと機械翻訳システムを直結すると、読取りと翻訳の間に次の2種類の編集が必要である。

(1) OCRの後編集(認識誤り文字の訂正)

(2) 機械翻訳の前編集(未知語検索、新語登録)

これらを独立した編集機能とするのではなく、一つの編集機能に統合すべきである。

3. 日本語ソフトOCR AS-Reader J

上述の機能を備えた機械翻訳システムASTRANSAC用日本語ソフトOCR(AS-Reader J)を開発した(図1、図2)。このOCRでは、複合類似度法を用いることにより99.5%以上の認識率を実現している[2]。

以下、機能の概要を述べる。

・レイアウト保存機能

AS-Reader Jはレイアウト自動認識機能[3]を有している。また、レイアウト認識の結果はMIF(Maker Interchange Format)形式に変換される。MIF形式はDTPツールAS-DocumentsおよびASTRANSACで採用しているドキュメント記述言語(レイアウト情報記録ツール)である。したがって、

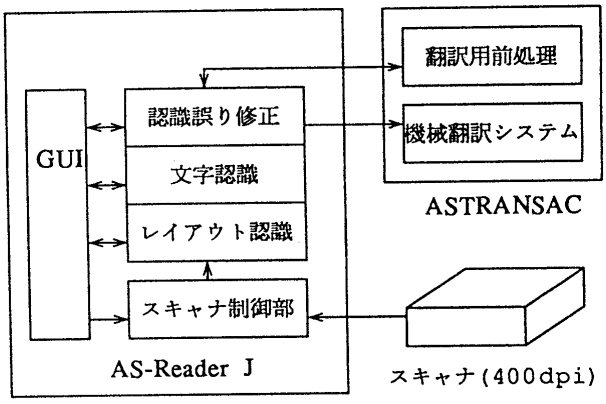

図1 AS-Reader Jの構成

Japanese OCR for Machine Translation System
Etsuo ITOH and Kimihito TAKEDA
TOSHIBA corp.

これらのシステムを読み取った原稿を翻訳した結果をレイアウト情報を用いて復元することができる。

また、縦書きの原稿を読み取った場合に、レイアウトを横書きに変換し翻訳結果の復元を行うこともできる。

・「文」の認識

従来のOCRでは以下の問題があった。

(1) 各行末に改行が付加されるため文が分断される

(2) 句点が読点として誤って認識されると、本来2文である文が接続される

これらの問題に対し、AS-Reader Jでは、レイアウト情報を利用し、パラグラフエンド以外の改行の削除、句読点の前後の品詞の解析による句読点の正答率向上という手法を取ることにより、「文」を正しく識別できるようにした。

・操作の一貫性

GUIにはOPEN LOOKを採用し、ASTRANSACとの操作性の統一を図っている。

また、AS-Reader Jで読み取った文書は、AS-Reader Jの制御画面から直接ASTRANSACに登録できる上、登録時のオプションで登録後翻訳、登録・翻訳後印刷の指示を行える。また、翻訳パラメータの設定も文書ごとに行うことができる。

・編集機能の統合

AS-Reader Jでは認識後の編集時に、認識誤りの検出をASTRANSACの形態素解析機能を用いて行い、認識誤りの訂正と翻訳の前編集のサポートを同時に行っている。また、編集中に機械翻訳用辞書への新語登録もできる。

4.おわりに

OCRをMTSの入力機能として捕らえ、機械翻訳、DTPと連動する事により、レイアウトを保存する翻訳システムを構築することができた。

今後はさらに必要な機能の検討を重ね、使いやすいシステムに発展させる予定である。

参考文献

[1] 伊藤、他 "DTP形式情報を保存する機械翻訳支援システム", 第42回情報処理学会全国大会論文集, pp.37-38 (1991)

[2] 有吉、他 "変形パターンの自動生成によるマルチフォント印刷漢字認識" 電子通信学会全国大会論文集, p.1465 (1987)

[3] S. Tsujimoto, et al."Understanding Multi-Articled Documents" Proc. 10th Int. Conf. Pattern Recognition, Atlantic City, New Jersey, pp.551-556(1990)

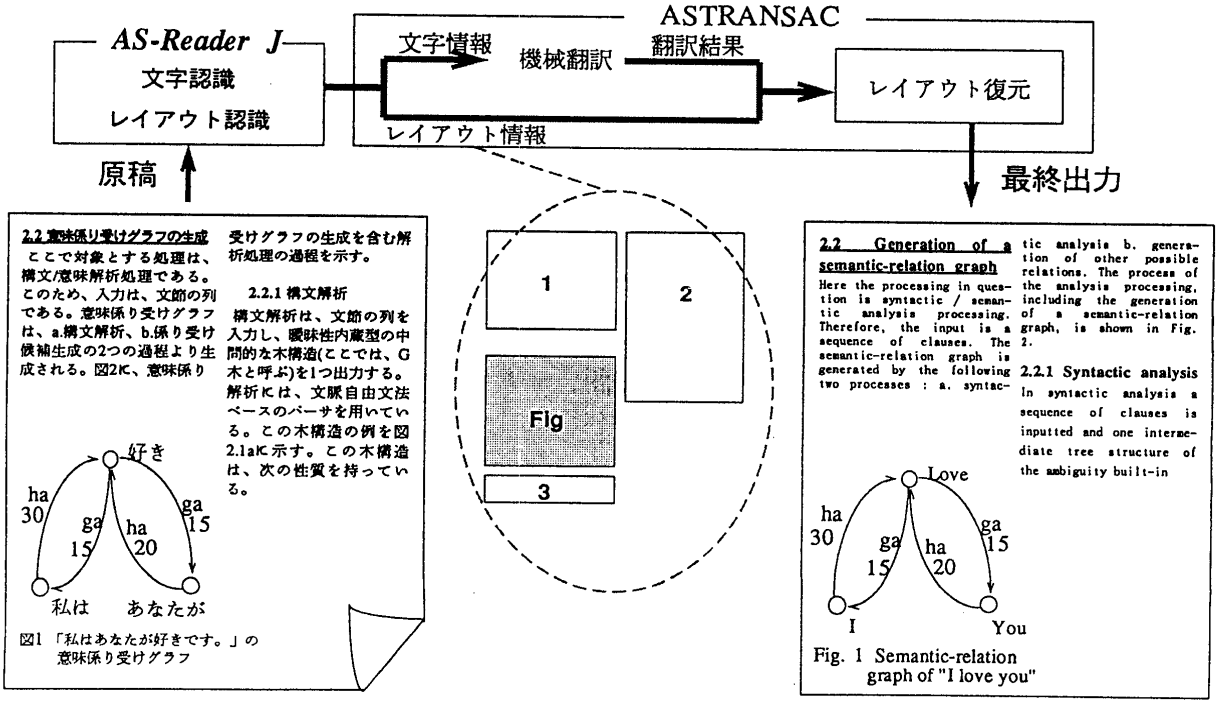

図1「私はあなたが好きです。」の意味係り受けグラフ

Fig. 1 Semantic-relation graph of "I love you"

図2 AS-Reader Jを用いた機械翻訳の流れ