

7M-7 WWW上での自動点訳サーバと未登録語の収集

生川武史 横平貫志 早川哲史 兵藤安昭 池田尚志

岐阜大学工学部

1 はじめに

我々は、現在、自動点字翻訳システムIBUKI-TENの開発を行っており、評価版をWWW上で公開している[1][2][3]。今回はさらに、IBUKI-TENをWWW上に自動点訳サーバとして構築することを試みた。これによって、インターネットに接続できるユーザは、点訳したい日本語テキストを準備するだけで、簡単な操作で点字化することができる。その際、ユーザ辞書を準備しておけば、それを使った点訳も可能である。また、点訳したいテキストに出現する未登録語を自動的に抽出してユーザに提示する機能もあり、ユーザは、それを参照しながらユーザ辞書を構築することができる。そのようなユーザ辞書を用いて再度点訳すれば、さらに高精度の点訳結果を得ることができる。

2 システムの概要

自動点訳サーバの構成を、図1に、利用画面を図2に示す。ユーザは、点訳したいテキストファイルをWebブラウザ上で選択し、Webサーバに送信する。送られたテキストデータをWebサーバで受信し、IBUKI-TENに渡す。IBUKI-TENでは、自動点訳を実行し点字コードを出力する。ユーザは、ブラウザ上から点字データをダウンロードすればよい。出力形式として、ひらがな分かち書き表記の点訳結果も選択することもできる。

ユーザが所有しているユーザ辞書をブラウザ上で指定すれば、IBUKI-TENは、サーバメモリ上のシ

ステム辞書にユーザ辞書を加えて点訳を行い、点訳が終了すると、ユーザ辞書は直ちに消去される。また、点訳したいテキストに出現する未登録語を収集し、ブラウザ上に表示する機能もある。現在のところ、漢字連続の未登録語を対象としている。ユーザは、これを基に単語登録を行い、再度点訳を実行すれば、より高精度な点訳結果を得ることができる。

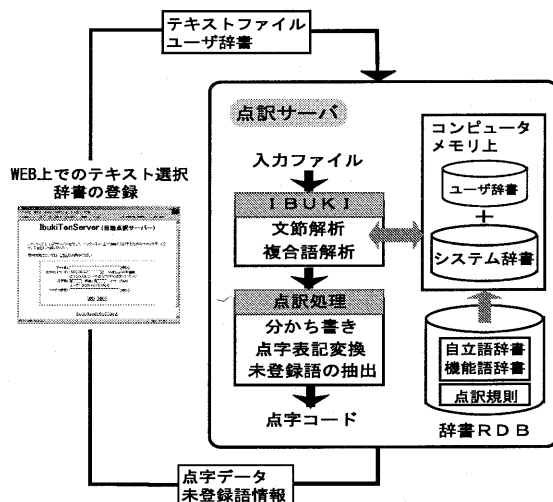


図1: WWW上での自動点訳の流れ

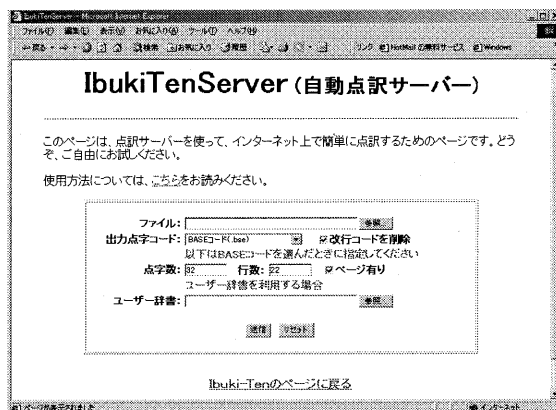


図2: 自動点訳サーバの利用画面

3 自動点訳サーバ

ユーザからブラウザ上で入力した情報が送信されると、Webサーバに用意されたCGI (Common Gateway Interface) によって、テキストファイルと点訳設定情報がWebサーバにアップロードされる。次に、CGIは、出力ファイル名を決め、ファイルの文字コードをsjisに変換する。出力ファイル名は、ユーザが指定したファイル名の拡張子を変更するようにした。そしてCGIは、IBUKI-TENに、これらの情報を渡し点訳を実行する。

IBUKI-TENでは、1文毎に、文節単位の切り出しを行い、漢字連続文字は、複合語解析によって名詞、接辞等に分割する。この文節解析の結果を基に、点訳処理では、点訳規則を記述した辞書を参照しながら、点字の規則に従った分かち書き、点字の表記法に従ったひらがな表記への変換を行い、点字コードを出力する。

IBUKI-TENは、Webサーバと非同期に起動させ、ブラウザには、「点訳中です」とのメッセージを表示する。その間、CGIは、定期的にサーバ上に出力ファイルが作成されていないかを確認する。点訳が終了して出力ファイルがWebサーバ上に作成されると、ブラウザに点字ファイルのダウンロード画面を表示し、ユーザはこのファイルをダウンロードすることで点字化されたファイルを取得できる。

4 ユーザ辞書と未登録語の抽出

IBUKI-TENでは、解析用辞書として主にEDR日本語単語辞書を用いて点訳を行っているが[4]、これで、あらゆる分野に対応できるわけではない。また、登録されている単語の読みや分かち書きを変更したい場合もある。そこで、ユーザが独自に、新たな単語を追加したり、登録されている単語を削除、変更してWeb上で点訳を行うことができるようにした。具体的には次のように行う。点訳を実行すると、点訳サーバがテキストに出現する漢字未登録語を自動的に抽出し、ブラウザ上に表示する(図3)。ユーザは、その単語の品詞、点訳表記を入力する。また、ダウンロードした点訳結果を見て、正しく点訳されていない単語があった場合は、ブラウザのフォーム上

に単語、品詞、点訳表記を入力する。そして、再度点訳を実行すれば、より高精度な点訳結果を得ることができる。ここで入力したユーザ辞書を点訳サーバからダウンロードして、各自で保存することで、次回からは、この辞書を指定して点訳が可能となる。



図 3: 未登録語表示画面

5 おわりに

現在開発中の自動点字翻訳システムIBUKI-TENをWWW上に自動点訳サーバとして構築することを試みた。今後、ユーザの意見を参考にしながら改良をしていきたい。また、ユーザの協力も得て、システム辞書の充実に反映させていきたい。

参考文献

- [1] 兵藤, 横平, 早川, 池田, 辞書データ主導型の自動点字翻訳システムIBUKI-TEN, 電子情報通信学会技術研究報告, WIT99-22, pp131-136, 1999.
- [2] 横平, 兵藤, 早川, 生川, 村上, 太田, 池田, 自動点字翻訳システムIBUKI-TENの校正支援機能, 電子情報通信学会技術研究報告, WIT00-18, pp37-42, 2000.
- [3] 自動点字翻訳編集システムIBUKI-TEN, <http://www.ikd.info.gifu-u.ac.jp/IBUKI-TEN/>.
- [4] 日本電子化辞書研究所, EDR電子化辞書仕様説明書, 1995.