4K－8 重要語抽出に基づく日本語マニュアルハイパーテキスト化ツール

内間 圭介　森 辰則　中川 裕志 *
横浜国立大学 工学部

1 はじめに

ハイパーテキスト化はマニュアルを読みやすくする手段の一つと考えられ、その作業の自動化に関する研究も発表されている[J.G96, 雨宮 96]。しかし、ハイパーリンクの自動生成はある程度の精度で行えるものの、現状では最終的に人手による確認をせざるを得ない。本稿では、この際の負担を軽減するものとしてタグ付け作業の支援をするツールについて述べる。本ツールではまず、ハイパーテキスト化する文書から索引語を抽出し、それをもとに機械的にリンクを生成する。その後 Web ブラウザを通じたユーザインタフェースにより文書のハイパーリンクの確認と訂正を行う。これにより複数文書間にわたるリンクを含むマニュアルのハイパーテキスト化を効率よく進めることが出来る。

2 システム概略

ハイパーテキスト化の作業は概略次の４ステップで行なわれる.

1. ハイパーリンクを張る語を選択する.
   オフラインで HTML 化する文書から索引語候補となる重要語を抽出
2. 選択された語を参照・被参照箇所に分類する.
   オフラインまたはブラウザ上で、利用者が指定した重要語抽出の結果ファイルと操作対象のファイルに対して定義パターン（後述）をもとに参照・被参照のタグ付けをする
3. 利用者による参照・被参照関係の編集
   ブラウザ上で索引語ごとにリンクのチェックを行なう
4. ハイパーリンク生成
   結果（HTML ファイル）をユーザに返す

ユーザはブラウザでドキュメントソースを保存することにより結果を受けとる。また、システム側には HTML 化する文書が中間タグを含んだまま残るので作業の中断や再編集には支障がない。次に各ステップの説明をする。

3 索引語の抽出

文書のハイパーテキスト化において、ハイパーリンクを張るべき語の抽出は最大の問題である。リンクを張るべき語はいわゆる索引語にほぼ相当し、マニュアルでは、これがマニュアルの主要な概念を示す名詞を含む複合名詞であることが多い。本ツールでは、我々が既に発表している索引語候補を重要語抽出システム[中川 96] により選び出す。この結果には各語の重要度などの情報が含まれるが、ここでは索引語候補を得るのが目的のため、重要度に関わらず抽出された語をすべて索引語候補とする。なお，本システムでは重要語抽出アルゴリズムとは独立であり，ハイパーリンクを張るべき語のリストを入力とするために，別の手法で生成された索引語を入力とすることもできる.

4 ハイパーリンクの生成

抽出された索引語候補を用いて、互いに関連しあう箇所を自動的に抽出することは可能であるが，リンクの方向，すなわち，参照・被参照関係を高精度で一意に決定するのは困難である．そこで，参照・被参照関係を推定し，これを初期値として参照関係エディタに渡す．利用者はその関係を必要に応じて修正することにより，正しいリンクを生成できる.

ただし、この時点で実際に文書に入れられるのは参照・被参照タグに相当する中間タグ（コメントタグ）であり、結果をユーザに返す際に対応する HTML タグに変換される。また、同一索引語に対するタグ付けは一段落ごとに一回ずつ行なわれる。

4.1 定義パターンの使用

マニュアルでは、索引語が見出し語または定義語として使われている箇所は被参照箇所になる事が多い。そこで、「（索引語）は～である」、「（索引語）を

*An Interactive Tool for Transrating Japanese Manuals into Hypertext
Keisuke Uchima. Tatsunori Mori and Hiroshi Nakagawa
Yokohama National University, 79-5 Tokiwadai, Hodogayaku, Yokohama 240,Japan

～と呼ぶ」といった定義パターンを集め、被参照箇所の特定に使用する。パターン集は実際にはPerlの正規表現に準拠したパターンを集めたファイルである。

タグ付けの際は、まず索引語を含む段落中でこのパターンに合うものを探し、見つからない場合は参照箇所とする。

4.2 参照箇所の間隔確保

実際に本システムのプロトタイプを利用して，得られた観察によると，おなじ語に対する参照箇所同士が近過ぎたり、被参照箇所のすぐ後ろにある参照箇所は修正段階でリンクの対象から外されることが多い。そこで、参照箇所同士および参照箇所と被参照箇所の間に一定の間隔を設けることにし、この間にある参照タグをリンクの対象から外す。実際の使用感から、デフォルトではこの間隔を5段落としている。この際、単に参照タグを取り外すのではなく、後で復帰可能な中間タグに置き換える。なお、被参照箇所についてはこのような間隔確保を行なわない。

5 Webブラウザによるリンクエディタ

生成されたハイパーリンクのチェックはWeb ブラウザで行なう。リンクのチェックは索引語候補ごとに行ない、チェックする索引語を選択するとタグ付けが行なわれた箇所が一文単位で表示されるようになっている。ここで、その内容を読んで、その部分を参照箇所とするか被参照箇所とするか、あるいはリンクの対象から外すかをラジオボタンの選択により決定する。ハイパーテキスト化する文書が複数ある場合でも同一ウィンドウ内で編集できる点がエディタでの作業と比較して有利である。

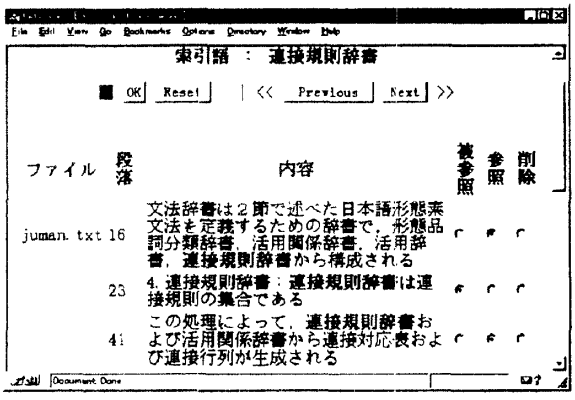

図 1: リンクエディタ画面

6 今後の課題

今後の課題としては、定義パターンの見直しが挙げられる。現在使用しているものは定義箇所になると思われるパターンを単に集めたものであり、このパターンの適用の際にも優先順位を設けていない。

しかし、本ツールの使用によりユーザの編集過程を入手できるので，定義パタンの正例と負例を収集できる．これらを用いてツールの利用を通じて漸進的に定義パタンを学習することも可能であろう．これは参照箇所の間隔の大きさについてもいえる。

7 おわりに

本ツールは、自動ハイパーテキスト化ではなく一部に人間によるチェックを加えたものであるが、6節で挙げた課題を検討することにより，システムの利用に応じて人手による訂正部分は減っていくものと思われる。またハイパーリンクの生成に使用する索引語候補は独自に追加することも可能であるため、OCR入力から得た索引語を使用するなど、信頼性の高い自動ハイパーテキスト化ツールの少ない現状では、実用面から見ても有用なツールであると言える。

8 謝辞

本研究はIPAの創造的ソフトウェアプロジェクトの援助をうけている。

参考文献

[J.G96] Stephen J.Green. Using lexical chains to build hypertext links in newspaper articles. AAAI 96 workshop on internet-based information systems, AAAI, 1996.

[雨宮 96] 雨宮秀文, 森辰則, 中川裕志. 重要語抽出による日本語マニュアルのハイパーテキスト化. 情報処理学会 第2回年次大会発表論文集, 情報処理学会, 3月 1996.

[中川 96] 中川裕志, 森辰則, 松崎知美. 日本語マニュアル文における名詞間の連接情報を用いたハイパーテキスト化のための索引語の抽出. 自然言語処理研究会報告 96-NL-116, 情報処理学会, 11月 1996.