定義パタンの学習に基づく自動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 電子情報工学科
E-Mail: {k-suke,mori}@forest.dnj.ynu.ac.jp, nakagawa@naklab.dnj.ynu.ac.jp

概要

マニュアルの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ている．しかし，相互参照情報 (ハイパーリンクあるいはリンク) の自動生成はある程度の精度で行なえるものの，現状では最終的に人手により確認・修正をせざるを得ない．そこで本稿では，これら人手による作業を軽減するマニュアルのハイパーテキスト化ツールについて述べる．このツールは，複合語に注目した重要語抽出システム，学習型のリンク生成システム，ならびに利用者向けのリンク編集システムから構成され，ほぼ自動的にハイパーテキストを生成することができる．学習精度は，テストに使用したマニュアルが2種類ではあるものの，交差検定による評価によれば90%程度の正当率で「定義箇所」，「参照箇所」の予測ができることが確認できた．

1　はじめに

本稿ではマニュアルのハイパーテキスト化を，ある語について，その説明がなされている出現箇所 (定義箇所) と，それ以外の出現箇所 (参照箇所) とを結びつける過程であると考える．このように，語に注目したハイパーテキスト化を前提にすると，自動ハイパーテキスト化に際して問題となるのは，以下の2点である．

1) リンクを張る語 (重要語) をどのように抽出するか
2) 抽出された重要語について，その出現箇所各々が，定義箇所であるのか，参照箇所であるのかというリンクの方向に関する判定をどのように行なうのか

まず，1) について述べる．リンクを張る語は書物の巻末にある索引語にほぼ相当するものである．ここでは，いわゆる重要語を考える．重要語抽出に関する研究は様々な観点からおこなわれているが，本システムでは，我々が既に発表している複合語に注目した重要語語抽出システムによって行なう．

つぎに，2) について述べる．リンクの方向の決定には，語を定義し，説明する際に現れる特有の言い回し，定義パタンに注目する．このパタンは，もちろん，人手で設定することも可能であるが，様々な文脈に対応するパタンを網羅するのは難しい．そこで，リンクの方向を利用者が簡単に編集できるエディタと，定義箇所における注目している語の周りの表層表現を学習するシステムを組み合わせることにより，人間の編集結果を訓練データとして漸近的に学習することで，より精度の高い自動ハイパーテキスト化を目指す．

2　システム概略

本システムにおけるハイパーテキスト化は図1に示す通り，以下の5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．
   HTML化する文書から索引語候補となる重要語を抽出する
2. 選択された語を定義箇所・参照箇所に分類する．
   抽出された索引語のリストと定義箇所を推定する定義パタン (後述) をもとに定義・参照のタグ付けをする
3. (利用者が) 専用エディタにより索引語にふられた定義・参照関係を後編集 (ポストエディット) する
4. 実際にハイパーリンクを生成する
5. 利用者の編集結果を学習し，定義パタンを更新する

同様にして，人手による編集と，その結果を用いた定義パタンの学習を繰り返すことによりリンクの決定精度を次第に向上させていくことができる．

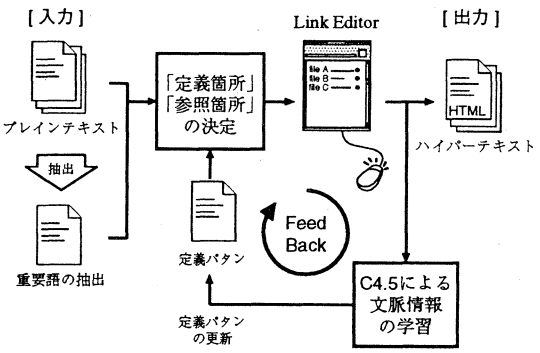

図 1: 本システムの概略

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重大な問題の一つである．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，索引語候補を，我々が既に発表している重要語抽出システム [中川 97] により選び出す。この重用語抽出システムは，5種類のマニュアルについての実験で，人間が見て索引語に選んだものを索引語の正解として，再現率 0.75，適合率 0.57 という評価を得ている．

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．

3.1　定義パタンの使用による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，ある語に対してどこを参照するかを高精度で一意に決定するのは困難である．

そこで，まず定義・参照関係を推定し，これを初期値としてリンク編集用エディタに渡す．利用者はこ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．本システムでは，編集結果を学習することによって，定義パタンを更新し，より人間の編集結果に近いリンクの方向の推定に使用する．

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が参照される事が多い．そこで，「(索引語) は～である」，「(索引語) を～と呼ぶ」といった，語を定義・説明する際に現れる特有の言い回しを人手で集めたものを，初期段階の定義パタンとして使用する．これは，Perl の正規表現に準拠したパタンを集めたファイルであり，85% 程度の正当率で「定義箇所」，「参照箇所」を推定できる．

4 Web ブラウザによるリンクエディタ

生成されたリンクの編集は，図 2に示すような WWW ブラウザ上のインタフェースを用いて行う．

編集は索引語候補ごとに行ない，リンクの編集対象とな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．ここで，その内容を読んで，その索引語についての定義箇所とするか参照箇所とするか，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．マウスクリックのみで複数のファイルにまたがるリンクを編集できる点が，通常のテキストエディタでの作業と比較して有利である．

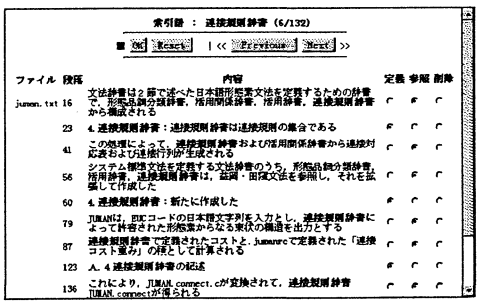

図 2: 編集画面の例

5 定義パタンの学習

3.1で述べたように，初期段階ではリンクの方向決定に使用する定義パタンは，定義箇所になると思われるパタンを人手により集めたものである．しかし，リンクエディタの編集結果から a) 索引語を含む文脈の表層表現 と b) その文脈に対する定義箇所 / 参照箇所の区別 (クラス) が得られるので，これを教師データとして索引語のまわりの表層表現を形態素解析することにより，a) と b) の間の関係，つまりここでいう定義パタンを学習することが可能であると予想される．

分類型の機械学習法においては，その方式だけでなく，どのような属性を，各学習事例から抽出すれば良いかという点が，学習精度向上にとって重要である．例えば，表層表現の情報だけで十分なのか，形態素解析の結果も必要なのかという点などは，漸進的な学習を考える場合，計算量の大小という観点からも興味深い．

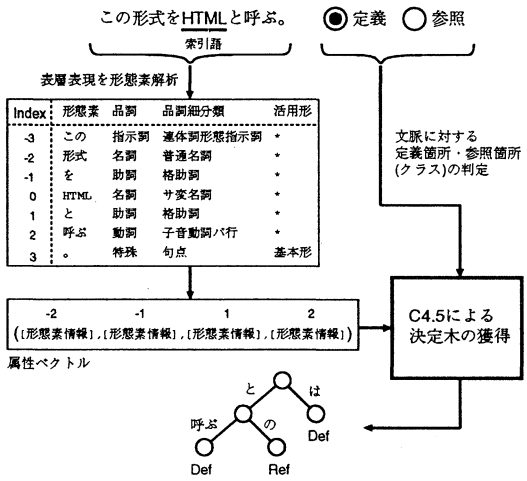

図 3: 定義パタンの学習

そこで，我々は編集結果から定義パタンを学習するための実験を行なった．概略を図 3に示す．この実験は，以下の情報を教師情報として，索引語の周囲の語 (文脈) からその索引語が「定義箇所」，「参照箇所」のいずれになるかを判定するための分類器を学習する．分類型の機械学習には様々なものがあるが，本実験では広く使用されていてその性質も良く知られている，決定木獲得アルゴリズム C4.5 を用いて決定木を得ることとする．

- 索引語を含む文の情報
  文を形態素解析し，索引語の前後一定数の形態素列について，それぞれ，「形態素基本形，品詞，活用形」の組をならべて一つの属性ベクトルを得る．この属性ベクトルは索引語が現れた周りの文脈情報を表す．
- その文が「定義箇所」，「参照箇所」のいずれになるかの判定
  人間により行なわれた編集結果から，上記の文それぞれについて，索引語を含むその文が「定義箇所」であるのか「参照箇所」であるのかの判定を得る．これは，上記属性ベクトルにより表される文脈において，索引語が「定義箇所」，「参照箇所」のいずれのクラスに分類されるかを表す．

実験では，表 1にあげる 2 種のマニュアルを使用した．表中で “平均形態素数” とあるのは，索引語を含む 1 文で，複合名詞を考慮したときの索引語前方・後方についての平均形態素数である．

なお，時間の都合上，JUMAN のマニュアルについては 3 人の被験者に編集をしてもらったが，ビデオデッキのマニュアルについては 1 人の被験者による編集結果を使用した．

また，ここでは学習精度のみに注目したいので，実験時に使用した索引語は，索引語抽出システムの出力に，人間による後編集を施している．これは

索引語抽出システムの適合率が100%ではないことと，リンクの編集および学習部分は，索引語抽出システムとは独立であるからである．

5.1 属性ベクトルとパラメタ

属性ベクトルの成分として使用する形態素情報，およびその数について，表2の組み合わせを考えた．属性値は形態素情報であるため，すべて離散値となる．ここで提案する属性ベクトルの構成法は，形態素の位置を，そのままベクトルの次元に対応させるナイーヴな方法であるから，形態素の認定ならびに，どの形態素を考慮するかによって，弁別精度が変わってくると予想される．よって，いくつかのパラメタを考慮して実験を行なった．

形態素解析機の名詞辞書項目への依存関係を減らし，実際の形態素に近付けるために，名詞列は複合名詞として1形態素と見なしている．“名詞を含む／含まない”というパラメタは，格助詞などの機能語と述語だけで，定義パタンを推定できるかどうかを知るためのものである．また，修飾語を取り除いた命題の骨格にすることにより，形態素の位置情報が正しくベクトルに反映されることが期待されるため，“形容詞・副詞を含む／含まない”というパラメタも導入した．

“表層表現を含まない”は，表層表現の多様さを品詞情報に捨象した場合の学習精度を知るためのパラメタである．“表層表現のみ”は，定義パタンの記述において品詞情報を使用しないものの方が，リンク方向推定時の，テキスト処理の際に計算量が少ないことから，表層表現のみを用いた場合の，弁別精度を推定するものである．

また，属性ベクトルとして使用した形態素列の，索引語に対する相対位置を表現するために，属性名に対して，索引語を原点(0)としたインデクスを付けている

5.2 「定義箇所」，「参照箇所」，「ヘッダ」の判定法

実際の実験では，索引語が節や項のタイトルなど，いわゆる見出し語として使用されている場合に，「ヘッダ」というクラスを新たに設け，被験者には「定義箇所」，「参照箇所」，「ヘッダ」の，3つのクラス分けをさせた．これは「ヘッダ」が，ハイパーテキスト化された時に読み手に参照されるという点では，「定義箇所」と同じであるが，まわりの文脈の性格が「定義箇所」とは異なるため，これを別のクラスにしておくことにより，より細かい分類が出来るであろうという予測に基づくものである．この実際の効果については後述する．

表1: 実験で使用したマニュアル

マニュアル		コンピュータソフトウェア JUMAN	家庭用ビデオデッキ 三菱電機 HV-F93
総文数		436	1461
大きさ(kB)		31	69
索引語数		133	190
索引語出現数		554	818
平均形態素数	前方	5.4	2.9
	後方	10.0	8.4

表2: パラメタの組み合わせ

属性ベクトルを作る範囲	索引語前後の形態素N個ずつ(計2N個)		
	索引語前方の形態素N個		
	索引語後方の形態素N個		
使用する形態素情報	表層表現＋品詞情報	名詞	含む／含まない
		形容詞，副詞	含む／含まない
	表層表現のみ	名詞	含む／含まない
		形容詞，副詞	含む／含まない
	品詞情報のみ	名詞	含む／含まない
		形容詞，副詞	含む／含まない

クラスの判定法としては，リンクエディタを用いて索引語を含む1文を提示し，次の方法でリンクの編集をさせた．

- 索引語についての定義・説明の記述があり，かつマニュアル読解支援のために，この箇所に至るハイパーリンクを張る必要があると思われる場合，これを「定義箇所」とする
- 同様に，索引語が見出し語として使われ，かつマニュアル読解支援のために，この箇所に至るハイパーリンクを張る必要があると思われる場合，これを「ヘッダ」とする
- 上記以外を「参照箇所」とする

6 実験の結果

表2に従って，諸パラメタを変化させながらC4.5により決定木を獲得し，交差検定による評価を行なった．属性ベクトルとして使用する形態素列の範囲，および，形態素情報について，各パラメタの組み合わせによる実験の結果を表3に示す．

使用した2種のマニュアルに対して，索引語を含む文が，「定義箇所」，「参照箇所」，「ヘッダ」のいずれになるかを判定するのに最適な，属性ベクトルの選び方の組合せは，次のようになった．

1. 形態素列は，表層表現と品詞情報を両方用い，名詞列を複合名詞として扱う
2. 名詞は残し，形容詞・副詞を除く
3. 属性ベクトルとして使用する形態素列の範囲は，索引語の左右に2個ずつ(計4個)とする

ここでいう“最適”とは，「獲得した木について，新規事例に対する予測誤り率が最小で，かつ，木の大きさがより小さくなるもの」とした．

1が示すことは，まず，表層表現だけではなく，品詞情報も用いることにより，文脈情報が適度に一般化されるということであろう．また，名詞列を複合名詞として扱うことにより，名詞列の長さによらず，文脈情報が適切に符合化されたと考えられる．

2については，形容詞・副詞を除くことで，属性ベクトル中の属性の位置と，文脈中の形態素の相対位値を適切に対応づけるとともに，属性値数が減る．また同様に，名詞を含むことで，索引語自体が複合名詞の一部である場合を，その複合名詞と別のクラスに分類できるため，誤り率が低くなった．また，今回獲得した木には現れなかったが，名詞を含むことでサ変名詞＋「する」(“定義する”など)を考慮することができる．

3については，左右に3個以上の形態素列を取っても，誤り率の改善はみられなかった．また，“索

表 3: 実験の結果

マニュアル	パラメタ				交差検定による新規事例に対する木の予測誤り率
	範囲 (形態素数)	形態素情報			
			名詞	形容詞・副詞	
JUMAN	索引語の左右に2個ずつ(計4個)	表層表現+品詞情報	含む	含まない	10.3%
		表層表現のみ	含む	含まない	11.0%
		品詞情報のみ	含む	含まない	10.6%
	索引語前方に2個の形態素				14.1%
	索引語後方に4個の形態素	表層表現+品詞情報	含む	含まない	12.8%
		表層表現のみ	含む	含まない	13.2%
		品詞情報のみ	含む	含まない	13.1%
ビデオ	索引語の左右に2個ずつ(計4個)	表層表現+品詞情報	含む	含まない	8.0%
		表層表現のみ	含む	含まない	10.6%
		品詞情報のみ	含む	含まない	8.0%

引語前方のみ", "索引語後方のみ" の形態素列を使用する場合についても調べてみた.

- 索引語前方の形態素列のみ使用する場合
  前方の形態素列だけでは, すべてを「参照箇所」にする木しか獲得できなかった. 表3の14.1% という予測誤り率は, この場合のものである.
- 索引語後方の形態素列のみ使用する場合
  形態素数は4個取る場合に最適となるが, 「索引後左右に2個 (計4個) の形態素列」を取る場合よりも弁別精度は2% ほど劣る.

これにより, 属性ベクトルとして使用する形態素列の範囲について, 「索引語前方に2個, 後方に4個の形態素列」を使用する組合せも考えられるが, 実際に決定木を生成して調べたところ, 予測誤り率の改善は見られなかった.

また, 実験では, 索引語に対する定義・説明が文として書かれている「定義箇所」と, 章見出しなどの「ヘッダ」を分けて考えたが, 「ヘッダ」を「定義箇所」に含めてクラスを2つにしても, 予測誤り率にほとんど変化はなかった. よって, リンクの編集の際に, 「定義箇所」と「ヘッダ」は同じように扱って良い.

7 関連研究

一般に, 自動ハイパーテキスト生成は, 1) リンクを張る対象をいかにして決めるか, 2) 決定した対象を文書中の関連部分とどのように関連付けるか, という2つの部分問題から構成される. これを踏まえて以下に自動ハイパーテキスト生成に関連する研究について述べる.

黒橋らは, 専門用語辞典を対象にハイパーテキスト生成を行なった. リンクを張るべき対象は, あらかじめ与えられている索引語と, 語句を定義する際の言い回しパタンをもとにテキストから抽出した語である. そして, 動議語関係などから作成したシソーラスや, カテゴリ分類 (人手も加わる) も用いる [黒橋 92].

中川らは, マニュアルの索引語の多くが複合語であるという事実に着目し, 重要語抽出を行なった.

雨宮らは, 重要語抽出によるマニュアルのハイパーテキスト化を行なった. ここでは, 黒橋らと同様に語句を定義する際の言い回しをもとに定義語を抽出し, リンクを生成している [雨宮 96].

我々のシステムは, 中川らのシステムと雨宮らのシステムを基礎に持つものである. 定義パタンに基づく点は, 黒橋らのシステムに同じであるが, その定義パタン自身を学習する点と, ハイパーリンクエディタとを連動した点が新しい.

8 まとめ

定義パタンの学習に基づく自動ハイパーテキスト化ツールについて述べた. その結果, テストに使用したマニュアルが2種類ではあるものの, 交差検定による評価によれば 90% 程度の正当率で「定義箇所」, 「参照箇所」の予測ができることが確認できた. 今後, 学習の対象とするマニュアルの種類・数を増やしていくことにより, より一般的な文書に対して有効な定義パタンの獲得が期待される. また, 定義パタンは複数持つことができるので, 文書の分野, および個人のニーズによる使い分けも可能である.

本システムは, 表層表現と品詞情報から, リンクの方向を決定するための, 定義パタンを獲得するものである. 今回は実験までに至らなかったが, 品詞を解析するシステムと, 語句を定義・説明する際に特有の言い回しがあれば, 他言語への応用も期待できる.

参考文献

[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. 言語処理学会第2回年次大会発表論文集, pp. 85–88. 言語処理学会, 3月 1996.

[黒橋 92] 黒橋禎夫, 長尾真, 佐藤理史, 村上雅彦. 専門用語の自動的ハイパーテキスト化の方法. 人工知能学会誌, Vol. 7, No. 2, pp. 336–345, 1992.

[中川 97] 中川裕志, 森辰則, 松崎知美, 川上大介. 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. 情処論, Vol. 38, No. 10, 1997.