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

雨宮秀文　森辰則　中川裕志
横浜国立大学 工学部 電子情報工学科

1 はじめに

最近、WINDOWSのHELP機能やWWWのホームページなどで見られるような、語句をクリックする事によってテキストからテキストへのリンクを張ることが出来るようなツールが増えてきた。このようなツールの利点は、ユーザが必要な知識－例えば理解できない語句の意味－といったような情報をマウスの簡単な操作によって容易に得る事が出来る点である。この利点をマニュアルに生かすと、そのマニュアル自体が非常にユーザに理解し易い便利なものになる。本論文

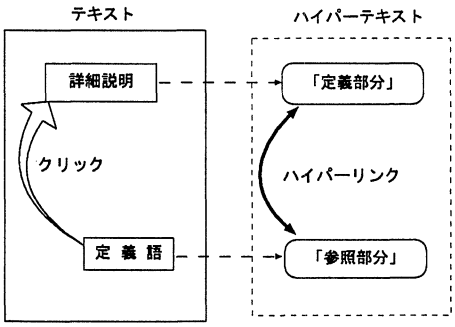

図 1: 本研究の目的

では、そのようなユーザが情報を必要とするであろう語句のうち、対象マニュアル内で定義されている語を自動抽出することと、それらへのリンクの張り方、また、それらの妥当性について重要度[和氣96]を利用する方法、等について述べる。

2 定義語とは

2.1 定義語の定義

マニュアル文においては特に顕著であるが、一般的に、ある語句を定義・説明した後、その語句を用いてまた新たな事柄を定義・説明するという構造をもった文書が多く見られる。
本論文では、このように文書内で定義・説明されている語句を**定義語**と呼び、定義語が実際に定義・説明されている文を**定義部分**と呼び、さらにその語句が使用されている部分を**参照部分**と呼ぶことにする。

1. 本体前面の右下のスイッチは電源スイッチです。
⋮
2. 電源スイッチをONの方にすると電源が入ります。

図 2: 定義語「電源スイッチ」の例

例を図2に示す。文1が定義部分、文2が参照部分となる。
本論文では、この定義語をキーとして「定義部分」と「参照部分」とをハイパーテキストのハイパーリンクという手法を用いて関係付けることを行なった。

このような事を行なうには、次のような手順で作業をすすめる必要がある。

1. 定義部分の候補となる文を探す。
2. 定義部分から定義語を切り出し、リストを作成する。
3. 文章中から各定義語の参照部分を探し、定義部分と結びつける。

2.2 定義語の出現パターン

さて、本論文での文中から定義部分と定義語を探し出す手法について述べる。それは、語句を定義する部分にはある特有の言い回しが存在するということである。したがって、定義部分で用いられるであろう表現を見つける事から始めた。その結果、具体的にいくつかのパターンを見つけた。これを、図3に示す。
従って、これらの定義部分に使われるであろう表現をもとに文章を調べていけば、このパターンに当てはまった文が定義部分の候補である、と言える。

- (内容) を (定義語) と定義する
- (内容) を (定義語) と呼ぶ
- (内容) を (定義語) と言う
- (定義語) とは (内容) である
- (定義語) は (内容) する

図 3: 具体的な定義パターンの一例

2.3 定義語の抽出

次に前の図3に示したようなパターンと一致した文、つまり、**定義部分の候補**から定義語を抽出する。その際、どのパターンとマッチしたかによって、定義部分における定義語の出現位置が決まっているので、定義部分から比較的容易に定義語が抽出できる。とは

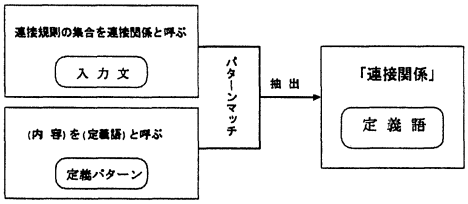

図 4: 定義語の抽出

いっても、このようにして抽出された語句が全て定義語である、とは言えない。そこで、最終的にこの語句が「定義語」となるかどうかを判定し、「定義語である」とされたならば、定義語としてリストに登録して、後の処理に使用する。現段階でのこの判断の基準は、

1. 指示語でない
2. 3章で述べる重要度を考慮する
3. 余計な語 (助詞等) が付属しない

といったものである。3. の場合には余計な語を削除したものが、定義語となる場合もある。

このようにして文章中から定義部分の独特なパターンに着目して、**定義語**を抜き出すのが本研究の目的の一つであった。本研究ではこれらを実現するのに、強力なパターンマッチング機能を持つ Perl 言語を用いて、システムを試作した。

定義語抽出の作業の後、本システムでは今作成した定義語リストを用いて文書内リンクを張るのであるが、これに関しては4章で述べる。

3 定義語と重要度

本章では、定義語を検索・判定する際に参考とする重要度 [和氣 96] と、実際に本システムで抽出した定義語との関係について述べる。

3.1 定義語と重要度との関係

まず、実際に「日本語形態素解析システムＪＵＭＡＮ」[JUM93] のマニュアルについて本システムを用いて抽出した結果と、その語句の重要度をプロットした結果を図5に示す。

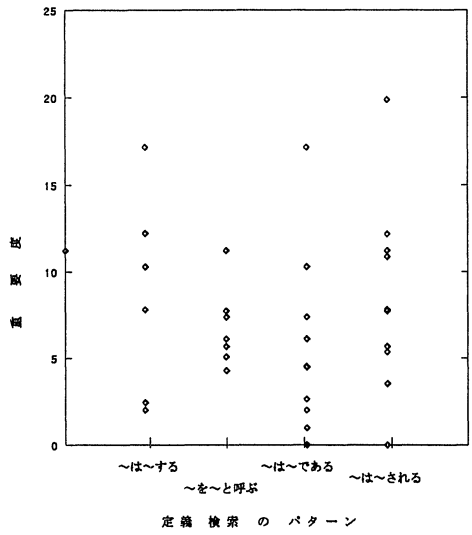

図 5: 「ＪＵＭＡＮ」の定義語と重要度のグラフ
縦軸は重要度、横軸が本システムで抽出されたかどうかを示している。

このグラフを見ると、本システムで抽出した定義語が、重要度においてほぼ一様に分布していることが分かるであろう。従って、両方のシステムが互いに補い合うことによって、両者ともさらに強力なシステムにすることが出来ると思われる。

3.2 定義語の重要度

本研究のような手法で抽出された定義語は、そのマニュアルで定義されているので重要な語句である。したがって、定義語の重要度は高い所に分布するはずで

ある。しかし、図5を見ると、重要度の高い所に集中して分布していない。

本システムで抽出された語句で、重要度の高くない語句は次のように分類される。

1. 名詞句 でない語句 である。
2. 連接せずに単独で用いられる語句 である。
3. 固有名詞 等である。
4. ゴミと思われるもの である。

このうち4. は重要度の値を用いて、削除できる。また、1. は、重要度の数値が求められないので、本システムの言語パターンのマッチングによって、重要語として抽出する。2. や3. は他の語と連接しにくい独特な概念の語句であろうと推測される。重要度は、その語句に連接する語句の種類の数をもとに求めたものなので、他の語句と連接しにくい語句の重要度は低くなってしまう。よって、本システムの言語パターンのマッチングによる結果を用いる。

3.3 重要度の高い語

重要度が高いにも関わらず本システムで抽出されない語は次のように分類される。

3.3.1 前提となる語句

マニュアルを読む以前に当然のこととして認識されていなければならない語句。
この例では、先の「日本語形態素解析システムＪＵＭＡＮ」のマニュアルにおいての「形態素」や、「形態素解析」といった語句であるとか、ビデオデッキのマニュアルの「ビデオ」「録画」「再生」といった語句が挙げられる。

3.3.2 一般語と連接した語句

一般的な接尾辞や接頭語が連接している語句。
この例では、「〜の例」とか、「〜の場合」といった一般的な語句が多くの語句に連接した場合においては、その語句の重要度が高くなってしまう。

同じランクの語句が並列的に多数用いられている場合は、「〜の例」といったような一般的な語句がその全てに連接して、結果として、「〜の例」という種類の語句の重要度が高くなってしまう場合が多い。

3.3.3 タイトルのみで出現する語句

タイトルのみで用いられいて、本文に出現しないような語句。

この例には、「〜の方法」といった節や章タイトルが多く用いられた場合がある。この場合、重要度は高くなるが、実際に本文中では使用されていない場合がある。

「〜の方法」といった節や章タイトルが用いられた場合、その節や章全体が その方法について述べられている。よって、その節や章にリンクを張ればよいがこのような場合には「〜の方法」という語句はあまり使用されていない。

いずれにせよ、重要度が高くて、本システムで抽出されない語は、本文中で定義されている場合が少なく、リンクさせるとしたら、外部のファイル等にリンクを張らなければならない語が多い、と言える。

4 定義語とリンク

本章では、本システムのうち、実際のリンクについてや、これまでに述べてきたような手順により抽出された「定義語」から、「定義部分」へリンクを張る作業について述べる。

4.1 文書内自動リンク

研究では実際に今まで述べた機能を既存のブラウザによって表示できるように**HTML**[1]のタグを用いた。HTMLのタグのうち、本システムで用いたタグはリンク関係の2種類である。定義部分に＜ＡＮＡＭＥ＝”ｉｄ”＞タグを用い、参照部分には＜ＡＨＲＥＦ＝”＃ｉｄ”＞タグを用いる。

今回は定義部分のタグは、定義文の1文全体をタグで囲むようにした。また、参照部分は登場した定義語の部分をタグで囲むようにした。タグを付けた出力例を図6に示す。

＜ＡＮＡＭＥ＝”１１”＞本体前面の右下のスイッチは電源スイッチです。＜／Ａ＞

⋮

＜ＡＨＲＥＦ＝”＃１１”＞電源スイッチ＜／Ａ＞をONの方にすると＜ＡＨＲＥＦ＝”＃１２”＞主電源＜／Ａ＞が入ります。

図 6: タグが付いた例

4.2 タグの付け方

本システムでのタグ付けの流れを示す。

[1]Hyper Text Mark-up Languageの略

4.2.1 定義部分のタグ

まず、定義部分のタグ付けであるが、これは、2章で述べたように定義語を抽出する際に、定義部分であると判断されて定義語が抽出された場合、その語句にid番号を付けて定義語リストに登録する。その時に定義部分にタグを付けて「定義部分にタグの付いたマニュアル」を作成する。これで、定義部分が全て抽出されて、その部分にタグが付いたマニュアルが出来た。

4.3 定義部分のタグ

次に、参照部分のタグ付けであるが、これは以下の手順で行なう。

1. 「定義部分にタグの付いたマニュアル」から一文読み込む。
2. 定義語のリストから一つ定義語を読み込む。
3. 入力文にその語が使われていたら、その語にタグを付ける。
4. 定義語のリストが終るまで2. と3. を繰り返す。
5. その文を出力し、マニュアルが終るまで1. から4. を繰り返す。

このような手順により、定義部分および参照部分の全てにタグを付ける事が出来た。

4.4 リンクの種類

リンクは2種類に分類される。図7に示したような**定義リンク**と**参照リンク**がある。これらについては、以下で述べる。

4.4.1 参照リンク

今までに述べたような 定義語 → 定義部分 というリンクの事である。定義部分をクリックする事で、その語の詳細説明が表示される。このようなリンクを「参照リンク」という。本システムでは、このリンクを自動で張るのが目的である。

4.4.2 定義リンク

一つの定義語についていくつか定義部分が存在する場合がある。例を挙げてみよう。「スイッチ」という定義語について2種類の定義部分、つまり、スイッチの場所について定義した部分と そのスイッチを何に使うかを説明した部分が存在する場合がある。これを 数珠つなぎ式 にリンクさせるのが「定義リンク」である。

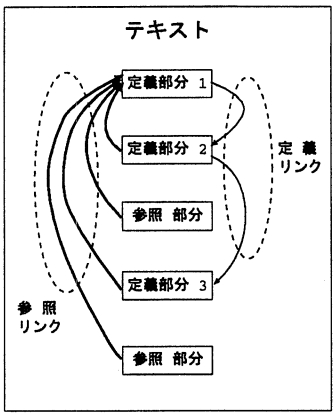

図 7: 2種類のリンク

以上、2種類のリンクにより、参照部分からその語句に関する定義を洩れなく辿っていく事が可能である。

5 おわりに

以上で述べたように、本研究では、
a) 「マニュアル中で定義された語句」＝「定義語」を抽出することが出来た。
b) 「定義語」をキーとして「定義部分」と「参照部分」の両者をリンクさせることが出来た。
c) 結果として,「ハイパーリンクの張られたマニュアル」への変換が出来た。

また、本研究は、3章で述べたように、重要語抽出の研究とも合わせて考えると、重要語の抽出といった分野でもそこそこ使用できるシステムであるといえる。

参考文献

[JUM93] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システムＪUMAN使用説明書 version1.0. 1993.

[和氣96] 和氣真 松崎知美 森辰則中川裕志. 語の連接の多様性に基づく日本語マニュアルからの重要語抽出. 言語処理学会第二回年次大会論文集, 1996.