文学データベースのためのプロトタイプシステムの実装

1V－2

本行弘明　白木善隆　田槙明子　國島丈生　横田一正
岡山県立大学 情報工学部

1　はじめに

ケルト文学の比較文学研究のために文学データベースシステムの研究開発を行なっている。総索引作成、文書の構造化、構造化文書の検索などの機能をもったプロトタイプシステムを実装した。本報告では、それらの中の文学データベース特有の機能と問題点を説明し、内容検索[1]や包括的な文書管理等の今後の研究開発方向を議論する。

2　文学データベースの構成

本研究における文学データベースは以下の3種類から構成されており、各々特有の検索機能がある。

- テキスト文書　- キーワード検索
  ケルト文学のディドラ伝説[2]の各伝本の本文を、ただ単に電子化した結果出来上がるフラットファイルのことで、一次情報である。
- 構造化文書　- 構造を意識したキーワード検索
  二次情報である。膨大な一次情報を調べやすいように章、段落、文などに構造化したものである。この構造化はXMLに類似したタグを挿入することによって行なわれており、構造情報は文章の特定の場所を特定するのに用いられる。
- ストーリ記述　- 内容検索
  因果関係など、詳細な関係が表現できているのが理想的であるが、それでは、必ずしも一意的な解釈が保証できない。また、ボトムアップでの記述だと物語全体を記述するのは非常に困難になる。そこで、本研究ではストーリ記述については基本的に
  - 完全なストーリ記述は不可能である
  - ストーリには多様な解釈が存在する
  - ストーリを場面間の有向グラフとして考える

  という考えに基づいて、物語のストーリを個々の場面に分割し時系列で結び、トップダウンでの記述にする。

3　プロトタイプシステム

今回実装したプロトタイプシステムは図1のように構造化、索引抽出、内容記述、検索の4つのサブシステムから構成されている。

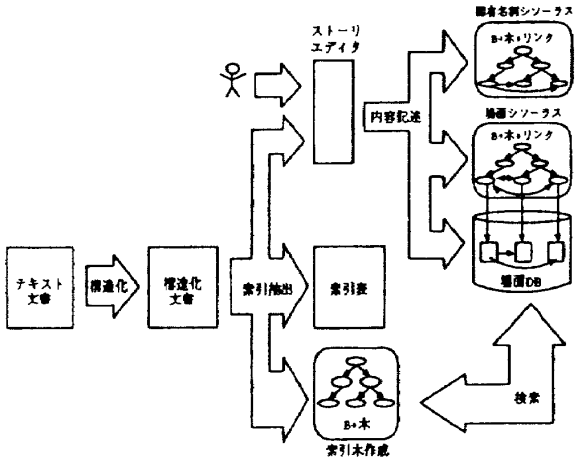

図1：プロトタイプシステム

3.1　構造化システム

一般的に文書の構造化は自然言語処理の構文解析に対応している。本研究での構造化の目的は前述したように

- 文書の位置の特定
- 検索の高度化

である。検索をより高度に行なうためには文章中の品詞の区別、指示代名詞の置換等を意味解析や文脈解析によって行なうことも考えられるが、本プロトタイプシステムでは実装とその評価を早急に行なう必要から文レベルまでの構造化にとどめている。

構造化の結果としてロケーション情報の利用が可能となる。図2の＜3,37,175,4＞といったものがロケーション情報である。この例でいうと、この文の位置は3章,37段落,175文,4韻文であるということを示している。

また、ディドラ伝説は文献によって文書の論理構造が異なり、構造化のための文法記述はある程度、それぞれの文書に依存している。また、ディドラ伝説の文章中には韻文（詩）で記述されている部分もあり、韻文には通常の構造化の法則が成り立たないため、構造化は困難であるということが問題点である。

3.2　索引システム

索引抽出は図1のように構造化文書に対して行なう。索引抽出の目的は

- 総索引表によって、用語の使用例や使用頻度を調べることで比較文学研究の支援ができる
- 索引木を作成することにより検索を高速に行なう

である。索引抽出は文書あるいは特定の部分に対しても行なうことができる。索引抽出を行なう際、抽出

Implementation of a Prototype System for Literature Databases.
Hiroaki HONGYO, Yoshitaka SHIRAKI, Akiko TAMAKI
Takeo KUNISHIMA, and Kazumasa YOKOTA
Faculty of Computer Science and System Engineering,
Okayama Prefectural University

語リスト、不要語リスト、固有名詞等の種類を指定することができる。さらに、統計情報の作成、出力形式(KWIC,KWOC)、B+木の作成の有無を指定できる。図2に総索引表の例を示す。各単語について、単語、その前後の文章、出現頻度、ロケーション情報で構成する。一番右端がロケーション情報で、数字は章、段落、文、韻文の番号である。

```
[explain]
1:   virion, Naoise, and explain it to me,' said Dei     <3,37,175,4>
[eye]
1:   off place, where no eye would see a sight o     <1,5,29>
2:   ote place, where no eye could see her and m     <1,6,33>
3:   have on before, no eye looked in her face     <2,9,45>
4:   he night is in mine eye;  I see Fearachar     <3,37,172,1>
5:   -hole and drove the eye out of the gay Geal     <4,48,223>
6:   though he drove the eye itself out of me, I     <4,48,229>
7:   her with the other eye had it not been for     <4,48,229>
[eyes]
1:   elf with his bodily eyes saw a blood-drop so     <2,23,117>
[face]
1:   o eye looked in her face but she instantly w     <2,9,45>
```

図2: 総索引表の一部

3.3 検索システム

3.3.1 全文検索

従来の全文検索と異なり構造情報が付加されているため、同一文章中、同一段落中といった条件を指定した検索が可能で、近接問合わせの拡張となっている。検索結果はロケーション情報であり、この内容は問合わせの内容に依存している。例えば、

- 同一文章中で $KW1 \wedge KW2$ が出現する文
- 同一段落中で $KW1 \wedge KW2$ が出現する段落

という問合わせに対しては、それぞれ文レベルと段落レベルのロケーション情報が返される。これは3.1で説明したように構造化が行なわれているためである。このような検索は索引レベルの操作で容易に実現できる。

文字列の照合としては、完全一致、前方一致については索引木を使用し、後方一致、中間一致などについては全文の照合を行なっている。また、必要に応じて逆索引木などを検討している。

3.3.2 内容の検索

内容記述は構造化文章、索引抽出などをもとにストーリエディタを用いて、人間の手で行なわれる。ストーリエディタとは、内容記述のデータ入力の負担を軽減するためのものである。内容記述の目的は、より内容に即した検索を行なうことと、複数の文書間の比較を容易にすることである。格納されるデータは、次のような要素で構成される。

- 固有名詞シソーラス: 構造化文章から固有名詞について索引木(B+木)を生成し、固有名詞同士関係あるものがリンクされる。つまり、B+木の葉の部分がリンクされる。
- 場面シソーラス: 索引システムによって、キーワードに対するロケーション情報が得られる。このロケーション情報と場面とを対応させ、キーワードに対する場面を検索する際に使用する。そして、場面の内容(場面データベース)にリンクされている。また、類似場面であるといった場面同士関係のあるものもリンクされる。
- 場面データベース: ストーリは場面を基にして記述されるが、場面を構成する要素は「場面ID」、「場面名」、「日時」、「場所」、「登場人物」、「キーワード」、「内容」であり、これらがデータとして格納される。個々の場面は時系列によりリンクされる。

内容検索の際には、このようにして作られた固有名詞シソーラス、場面データベース、場面シソーラスを対象として検索を行なう。

固有名詞シソーラスを利用して登場人物名を検索することができる。例えば、"Deirdier" をキーワードとして、検索を行なうと、上位語として "woman" が、同義語として、時代とともにスペルが変化してきた数々の "Deirdier" を指す単語などが、そして、下位語として、"Deirdier" の派生語などが検索結果として出力される。

場面の検索は場面シソーラスを利用する。あるキーワードとなる単語を入力すると、場面シソーラスから、その単語に関係のある場面の候補の一覧が出力される。そして、その中から、任意の場面を選ぶことで、場面データベースから、選ばれた場面の内容データが出力され、場面シソーラスからは、その場面に関係のある場面(類似場面、次場面)の一覧が出力される。

4 おわりに

現在は1つの文献についてのみの索引や検索しかできないが、複数の文献に対しても索引や検索を行うことを可能にし、比較文学研究の効率向上を図ることを計画している。

謝辞

種々の議論を頂きました横田研究室の皆様および岡山理科大学劉助教授に感謝致します。なお、本研究の一部は文部省科学研究費(基盤研究C)によるものである。

参考文献

[1] 本行弘明, 池口仁誠, 三宅忠明, 横田一正: 分散環境での文学データベースの内容検索, 情報処理学会第55回全国大会講演論文集(3), pp. 290-291 (1997).

[2] 三宅忠明編著: Select Versions of DEIRDIRE, 大学教育出版,(1998).