電子化国語辞書を用いた概念ベース自動構築における前処理の自動化

1M－5

小島 一秀　渡部 広一　河岡 司
同志社大学大学院 工学研究科

1 はじめに

知的な情報処理システムにおいては，人間らしい連想が重要な役割を果たす．人間らしい連想を行う連想メカニズムには，単語とその関係情報を格納する概念ベースが必要となる．この概念ベースの情報を十分に用意するためには，電子化国語辞書（以下“国語辞書”）などから自動的に作成する必要がある．

国語辞書を用いた概念ベース自動構築では，概念ベースに情報を格納する前に，国語辞書の文字列を概念ベース向け情報へ変換する前処理が必要となる．概念ベース向け情報とは，連想元単語と連想単語の対と，これら単語間の関係情報の組である．本稿では，この国語辞書の文字列を概念ベース向け情報に変換する前処理方式を提案する．

2 国語辞書からの単語の切り出し

提案する前処理方式は，辞書の表記構造を利用している．単純に説明文中の単語を見出し語の連想語とすると，同じ綴りで意味の異なる単語が区別されないなどの問題がある．例えば，“本”（表1）ならば，読む本，助数詞の本，“正式な”と言う意味の接頭辞などが区別されない．このように取得した情報で連想を行うと，例えば“本”からなら，表１の単語が全て出力される．

表1：単純な処理で得た“本”の各項目からの連想語
（自然言語部分の自立語取り出し）

読む“本”の項目	助数詞“本”の項目	接頭辞“本”の項目
書物，自分，書く，小説，本，する，本，シテ，出版，スル，書籍，図書，書，書冊，冊子，…	棒，細長い，もの，数える，語，マッチ，一本，映画，作品，数，数える，語，…	主な，正式，本舞台，…

3 国語辞書の表記構造

国語辞書の説明部は，二種類の要素がある規則的な構造をなして並んでいる．二種類ある要素の一つは文要素である．これは，自然言語部分の中の一部分で，連続した自然言語の文字からなる．もう一つは，記号要素である．これは，国語辞書特有な記号で自然言語にはない．規則的な構造とは，ある特定の記号要素によって定義された木構造の中に，文要素とそれを修飾する記号要素が並んだ構造である（図 1）．

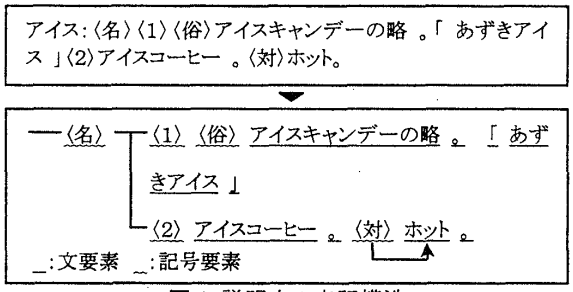

図 1：説明文の表記構造

4 概念ベース構築の前処理

4.1 概念ベース構築の前処理概要

国語辞書の説明文の文字列を概念ベース向け情報へ変換する前処理方式の基本方針は次のようになっている．原則的には，見出し語とその説明文の単語とは関係があるとみなして情報を取る．しかし，この処理は単純には行わず，次のような処理で行う（図 2）．まず，コンパイラ的な処理で説明文の文字列を，表記構造から得た情報と自然言語の組に変換する．そして，この組の自然言語を形態素解析と単純な処理で，自然言語から得た情報，単語の組に変換する．この組と表記構造から得た情報を合わせて概念ベース向け情報とする．

したがって，前処理は大きく分けて二段階の処理である．一段階目を構造処理，二段階目を自然言語内処理と呼ぶ．それぞれの詳細を以下に述べる．

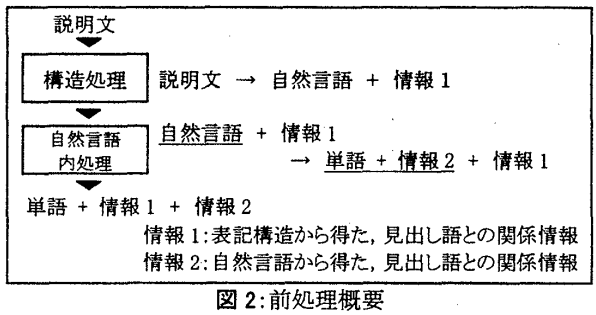

図 2：前処理概要

4.2 構造処理

構造処理は，説明文の規則的な表記構造に対する処理により，説明文の文字列を単位文，分類番号，見出し語との関係情報の組に変換する．単位文とは説明

Preprocessor for Automatic Concept Base Construction with Electronic Language Dictionary
Kazuhide KOJIMA, Hirokazu WATABE, Tsukasa KAWAOKA
Graduate School of Engineering, Doshisha University

文中の自然言語文字列を"。"で区切った文字列である．構造処理は，要素分解，項目展開，文展開の三段階からなり次のようになっている(図3)．

要素分解： 説明文の文字列を，文要素と記号要素に分解する．コンパイラの字句解析[1]に相当する．

項目展開： 説明文を項目ごとに分解する．同時に，各項目に項目番号を割り当てる．

文展開： 項目ごとに分解された文字列をさらに，単位文に分解する．このとき，記号要素によって付加される見出し語との関係情報を付加し，不要な要素を削除する．

入力：図1（アイスの説明文）

構造処理

要素分解：文字列を要素に分解する．"|"は切れ目．

〈名〉|〈1〉|〈俗〉|アイスキャンデーの略|。|「|あずきアイス|」|〈2〉|アイスコーヒー|。|〈対〉|ホット|。

項目展開：項目で展開する．項目を示す要素を削除．

分類	項目の内容
1	〈名〉~~〈1〉~~〈俗〉アイスキャンデーの略。「あずきアイス」
2	〈名〉~~〈2〉~~アイスコーヒー。〈対〉ホット。

文展開：文要素で展開する．処理に使えない要素を削除．

分類	関係	単位文
1	—	~~〈名〉〈俗〉~~アイスキャンデーの略。
1	—	「あずきアイス」
2	—	~~〈名〉〈2〉~~アイスコーヒー。
2	対	~~〈対〉~~ホット。

出力

分類	関係	単位文
1	—	アイスキャンデーの略。
1	—	「あずきアイス」
2	—	アイスコーヒー。
2	対	ホット。

図3：構造処理の例

4.3 自然言語内処理

自然言語内処理は，構造処理の結果である単位文を，単語及び，見出し語との関係情報に変換する．これは，形態素解析，定型単位文－国語辞書で頻出する定型的な単位文－を利用した関係情報取得と不要単語の削除による．形態素解析には茶筌[2]を利用している．

定型単位文を処理する方法は次のようになっている．文字列一致で定型単位文を検出し，不要な文字列を削除する．そして，形態素解析後に場合により特定の位置にある単語の関係情報を取得する．文字列一致のチェックは形態素解析前に，関係情報の取得は形態素解析後に行う必要がある．この形態素解析前の処理が定型文検出，形態素解析後の処理が形態素検出である．まとめると以下のようになる(図4)．

定型文検出： 定型単位文を検出し，不要文字列を削除する．

形態素解析： 単位文を品詞に分解する．自立語以外を削除する．

形態素検出： 特定の品詞配列を検出し，形態素解析で出力された単語と見出し語との関係情報を取得する．

表2：定型単位文を処理するルールの例

条件	処理
"…の略。"	"の略。"直前の単語は見出し語と同義語．"の略。"を削除．
"の意を表す。"	"の意を表す。"を削除．

入力：構造解析（図3）の出力

自然言語内処理

定型文検出：定型単位文の検出．定型部の削除．

分類	関係	定型文	単位文
1	—	の略。	アイスキャンデー~~の略。~~
1	—	—	「あずきアイス」
2	—	—	アイスコーヒー。
2	対	—	ホット。

形態素解析：単語に分解．付属語と記号を削除．"|"は切れ目．

分類	関係	定型文	単位文
1	—	の略。	アイスキャンデー
1	—	—	~~「~~|あずきアイス|~~」~~
2	—	—	アイスコーヒー|~~。~~
2	対	—	ホット|~~。~~

形態素検出：品詞の配列から関係情報取得．

分類	関係	定型文	単語
1	同	の略。	アイスキャンデー
1	—	—	あずきアイス
2	—	—	アイスコーヒー
2	対	—	ホット

出力：概念ベース向け情報

分類	関係	連想単語
1	同	アイスキャンデー
1	—	あずきアイス
2	—	アイスコーヒー
2	対	ホット

図4：自然言語内処理の例

5 おわりに

提案した前処理により，国語辞書から概念ベースを自動構築する際に，自然言語以外の除去，同表記異義語の区別，単語間関係情報の取得が可能となった．

この処理方式の持つ課題としては，誤った単語の取得，見出し語の分類の過度な細分などがあげられる．また，前処理の結果を概念ベースに反映させる方法も今後の課題である．

参考文献

[1]佐々政孝：プログラミング言語処理系，岩波書店(1989)

[2]http://cactus.aist-nara.ac.jp/lab/nlt/chasen.html