E-15 意味構造抽出によるテキストマイニングと知識利用
Text Mining and knowledge Utilization by Extracting Semantic Structure

中原 大介† 杉浦 優介‡ 岸 義樹†
Daisuke Nakahara Yusuke Sugiura Yoshiki kishi

1. 概要

本研究は、日本語のテキストデータを解析し、その文章構造を把握した上で、テキストデータの表す意味の知識ベース化、更にはその知識ベース利用による知識発見を目的とする。その方法として、まずはテキストデータに対して自然言語処理を行い、文章構造を抽出する。ここでは、文を単位に解析を行い、文を単文の集合として表現し、それをもって文の構造とする。また、単文の意味解析については格文法を用いて、格フレームで表現する。

文章は文の集合であるため、文フレームを利用して接続詞などの品詞情報を基に文章の知識表現を行う。また、文や文章のフレームにはユニークな識別子を付加し、参照の効率化を図る。以上のような手順によりテキストデータの知識表現を行い、知識ベースを構築する。

本研究は文章の意味解析と知識ベース化という、自然言語処理と文の意味構造を含んだ知識を利用している点に特徴がある。また、本研究で作成したシステムは特定分野に特化しておらず、融通の利くシステムであることにも利点がある。

2. システム構成

本研究でのシステムは、大きく分けて2つの部分から成る。1つはテキストデータ中の品詞出現頻度や文法的特徴をとらえ、テキストに前処理を施し、次のステップである文章構造解析にデータを受け渡すためのテキスト整形システムである。もう1つはテキストデータから文章の意味構造を抽出するためのテキスト解析システムである。なお、使用言語は主に K-Prolog である。

2.1 テキスト整形システム

対象とするテキストデータには一定の形式があるわけではなく、多くのノイズを含んでいる。テキスト解析をスムーズに行うために、テキストデータの前処理が不可欠となる。

本システムではテキストデータの前処理を2段階に分けて行う。パーザによる形態素・構文解析に支障があると思われるデータをテキストデータから取り除く前処理と、形態素・構文解析から、文章構造抽出に支障のある形態素・文節情報を削除・訂正する構文情報処理のステップである。前処理の流れを図1に示す。

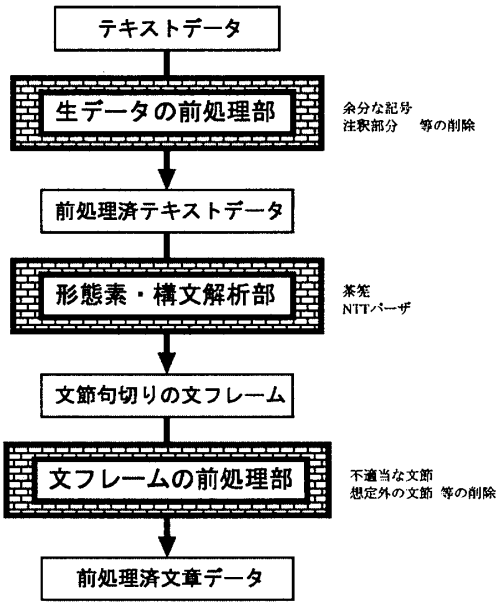

図 1: テキスト整形システムの処理の流れ

1. 前処理
テキストデータから解析の妨げとなるノイズを処理する。主に、記号の除去、半角文字から全角文字への置換、括弧で括られている部分の除去を行う。同時に、テキストデータから見出し等を参考にして文章データを取り出し、1行1文の形式でファイルに格納する。

2. 形態素・構文解析
構文解析は文を文節に区切るのみに留める。解析には NTT の日本語形態素解析パッケージ ALTJAWS Ver.2 を主に使用している。解析結果は以下のような Prolog 項の形式で表現する。

```
文：sentence( 文番号, 文節リスト ).
文節：clause( 形態素リスト ).
形態素：morph( 見出し, 品詞情報コード,
               語彙大系意味属性 ).
```

3. 文フレームの前処理
形態素・構文解析結果における文節フレームの形態素並びをチェックし、テキスト解析システムでノイズとなりうる文節フレームを削除する。削除は情報欠落を防ぐために最小限に留める。

2.2 テキスト解析システム

以下の3つの部品から構成される。

- 文型判断部
  入力文のタイプを特定し、文の妥当性を検証するため、単文/複文の特定、単文の非文/適格文検証、分割文の処理の3つの処理を主に行う。
- 複文分割部
  複文は、合文・重文・有属文の3タイプに分類して扱う。登録されたルールを元に各々を2つの文に切り分け、文の分割、主格の補完、複文フレームの作成の3つの処理を主に行う。
- 単文解析部
  単文を解析し、格文法に基づいた格フレーム表現を獲得する。格フレームは以下のような形式である。

```
sentence( 文番号,[ [ 名詞節リスト ],
                   [ 副詞リスト ],
```

†茨城大学工学部
‡日立ソフトウェアエンジニアリング株式会社

[述語節]]).
名詞節：[主名詞，助詞，
副名詞リスト，主名詞カテゴリ]
述語節：[述語，様態リスト]

この3つの部品が図2に示すような流れで文の解析を行う。

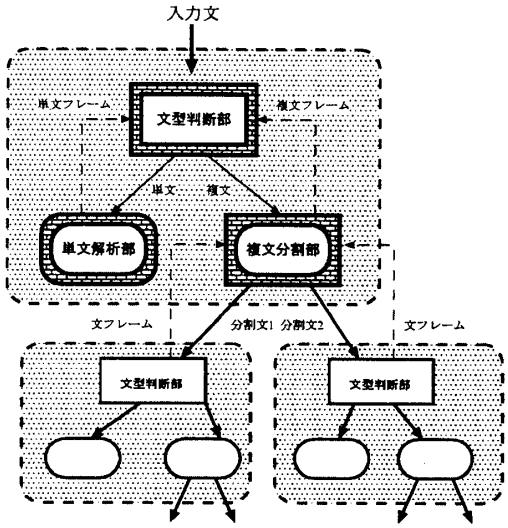

図 2: テキスト解析システムの処理イメージ

3. 知識ベースの構築と利用例

3.1 知識ベースの構築

本研究で使用したデータは「中小企業白書2001年版」[3]、「平成13年度 情報通信白書」[4]の事例文である。テキスト整形システムにより文の前処理を行い、その結果得られた特徴を表1に示す。

表 1: 事例文章のデータ

事例数	116	
文数	1146	1事例あたり約9.9文
文節数	16550	1文あたり約14.4文節
形態素数	36751	1文あたり約32.1形態素
名詞節	10811	全文節の65.4%
述語節	4709	全文節の28.4%
副詞節	531	全文節の3.2%
その他の節	499	全文節の3.0%

前処理されたテキストデータをテキスト解析システムに通して、知識ベースの構築を試みたところ、解析の成功した文は、事例文章中の1146の文の内の843文であった。これは、解析文の73.6%に当たる。

3.2 知識ベース利用例

構築した知識ベースの評価と利用例として、テキスト検索を行った。検索対象は、先の実験で用いた116の事例文とする。

- キーワード検索
  キーワードとして「ネットワーク」を与えた場合、単純検索では17の文章が結果として得られた。知識ベース利用による検索では、これがさらに5事例に絞り込まれた。ここで、この5事例は全てネットワークという語が話題として用いられている文を含んでいた。つまり、単純検索よりもキーワードの重みの強い事例が選ばれたことになる。

- フレーム利用による知識検索/発見
  知識ベースを利用した知識検索/発見の例として、事例文中から「開発したもの」を対象オブジェクトとして検索する例を示す。開発するものは「～を開発する。」というように「を格」で示される。従って、キーワードに「開発する」という動詞を与え、これとマッチする述語節を探し出し、その格フレームを参照して「を格」の名詞節を抽出することで、何を開発したのかを得ることができた。解析結果には抽象的なものが多く含まれるが、より具体的な情報を得るためにはその前後の文の参照も必要となる。

  本研究の手法を使えば、知りたい情報を直接検索できるという大きな利点がある。又、「で格」や「が格」を検索することで、何を用いて開発したのかや、誰(何)が開発したのかが検索でき、幅広い応用が可能となる。これらを複数組み合わせることや、キーワード検索との組み合わせによる解の絞り込みも可能である。

4. 終わりに

本研究において、日本語文章を格文法に基づく格フレームにより表現し、知識ベースを構築するシステムを作成した。作成したシステムにより、意味の絞り込みが有効であることを確認でき、格フレームの構造を利用して知識の検索/発見が可能であることも示された。

一方で、解析と知識ベースの構築に時間を要すること、テキストデータ中に存在するノイズに弱い等の問題が今後の検討課題である。

参考文献

[1] CD-ROM『日本語語彙大系』: NTT コミュニケーション科学基礎研究所 (1999)

[2] 日本語形態素解析システム『茶筅』version2.2.8 使用説明書: 奈良先端科学技術大学院大学 松本研究室 (2001)

[3] 中小企業庁 編: 中小企業白書2001年版: ぎょうせい (2001)

[4] 総務省 編: 平成13年版 情報通信白書: ぎょうせい (2001)

[5] Charles J Fillmore 著, 田中春美, 船域道雄 訳: 格文法の原理: 三省堂 (1975)

[6] 杉浦優介: 意味構造抽出によるテキストマイニングと知識発見: 平成13年度 茨城大学大学院理工学研究科 情報工学専攻 修士学位論文

[7] アイザック: K-Prolog Compiler User's Manual