RWCにおける品詞情報付きテキストデータベースの作成

井佐原 均（電子技術総合研究所）　元吉 文男（電子技術総合研究所）
徳永 健伸（東京工業大学）　橋本 三奈子（情報処理振興事業協会）
荻野 紫穂（日本アイ・ビー・エム株式会社）　豊浦 潤（新情報処理開発機構）
岡 隆一（新情報処理開発機構）

1 はじめに

RWCP（新情報処理開発機構、Real World Computer Pertnership）では、平成6年度よりRWCデータベースワークショップを設置し、実世界に関するデータの収集と利用を目的として、テキスト・音声・画像からなるデータベースの作成を行なっている。また、これら3分野の融合として、マルチモーダルデータベースの作成を計画している。

実世界（Real World）における自然言語処理の研究を行なうためには、電子化されたテキストデータベースが必要不可欠である。また、技術の健全な発展と評価のためには、このようなテキストデータベースは広く公開され、研究者が同一のデータに対して実験を行なえることが必要である。

欧米では既にこのような研究・評価用のテキストデータベースが作成され、研究者や企業に対して、共有資源として広く公開されているが、日本においては、この種のデータの組織的な作成は未だ十分には行なわれていない。また、入手可能なデータは高価な場合が多い。

このような点を踏まえて、RWCデータベースワークショップのテキストグループでは、研究・評価用に公開することを前提として、言語情報を付加した現代日本語のテキストデータベースを作成することとした。

2 RWCテキストデータベースの基本的立場

RWCテキストデータベースは以下の条件を満たすことを目標に作成している。

1. 大規模であること
2. 現実のテキストを反映したbalanced corpusであること
3. 精密かつ正確な情報を付加したテキストデータベースであること
4. 必要に応じて対訳テキストデータベースについても検討すること

また、作成の基本理念は「研究・評価を目的としての無償公開」「作成時の協調・分散」「言語理論から独立した汎用性」である。

「公開」については学術目的であれば、誰もが使える共通の資源とすることがもっとも大切な条件であると考えた。ただし、著作権の問題がこの種の言語データ共有においては、常に問題となる。今年度は既に他機関が公開を行なっている、あるいは著作権を主張しない、ものについて、RWCとしての加工と公開を行なうこととした。来年度以降はRWCとして積極的に著作権者と交渉し、公開できるデータを増やしていく予定である。

なお、現段階では、完成したテキストデータベースの配布方法については、確定した案はない。小規模な範囲でのテスト利用からのフィードバックにより、テキストデータの収集・加工についての示唆を得ることから順次始めていく予定である。

「協調・分散」とは、テキストデータベースの共有化を目指す他組織と相互に連絡を取り、協調しながら、ただしお互いは独立して、テキストデータベースの作成を行なおうというものである。具体的には、ＩＰＡコーパス［1］を作成中の情報処理振興事業協会とは、言語情報の付与に用いる品詞体系を共通のものとし、また、互いに収集するテキストの重複を避け、全体としてバランスの取れた言語データを集積するように注意している。（社）日本電子工業振興協会（電子協）においては、その年度報告書のテキストデータベース化および共有化を進めている［2］が、これは未加工テキストであり、ＲＷＣではその未加工テキストを得て、品詞情報の付加を行なっている。このように分野、形態（未加工データと加工データ）、加工作業（品詞体系）のそれぞれについて、「協調・分散」を実現している。

「汎用性」の考え方は後述する品詞体系の設定にもっとも強く反映している。ここでは特定の言語理論に依存するのではなく、出来るだけ多くの情報を付与しておくことにより、利用者がどのような理論に基づく品詞体系を用いようとしている場合にも（比較的）容易に変換できるように品詞体系を作成した。

3　ＲＷＣテキストデータベースの概要

テキストデータベースは、未加工テキストを集めたものと、言語情報を付加したものとに大別される。今年度ＲＷＣで作成したものは、言語情報を付加したものである。

言語情報の付加としては、（1）単語分割、（2）各単語への読みや品詞の付加、（3）係受け構造等の構文情報の付加、（4）意味情報の付加、等が考えられるが、今年度は、（1）および（2）を対象とした。来年度以降は、これに加えて（3）の係受け情報を加えたデータベースの作成を予定しており、今年度はその書式の検討を行なった。

3.1　対象とするテキスト

ＲＷＣでは、今後も継続的にテキストデータの収集・加工を行なうことにより、このテキストデータベースを balanced corpus としていく予定である。その一部をなすものとして、今年度は主として「公開」の可能性に注目して対象テキストを選択した。

今年度、対象としたテキストは、以下に示す１１１７９文からなる。

通商白書（編集・発行 通商産業省）
　　平成４年度版（２３７７文）　　平成５年度版（３２１４文）　　平成６年度版（１７７１文）
１９９４年版 我が国産業の現状－図とデータでみる産業動向（通商産業大臣官房調査統計部 編）
　　マクロ編（２８７文）
電子協平成４年度機械翻訳システム調査委員会報告書「機械翻訳システムの実用化に関する調査研究」
　　（３５３０文）

これらのテキストはよくこなれたものであり、自然言語処理研究用のテキストとしては適切なものである。通商白書の一文は比較的長く、本文では平均で６０文字弱である。電子協の報告書ではこれよりも多少短く、４０文字強である。この値は概ね、新聞の社説と同等である。

3.2 品詞体系

ここでは、品詞体系作成の基本方針として、「汎用性」を考えた。単語に品詞を付与する場合には、ある程度主観的な判断が必要となる。品詞体系そのものについても、各自の立場によって、受け入れ難い場合もあろう。ここでは、利用者が自分の研究目的に合わせて（自由に）取捨選択あるいは変更して利用できるように品詞体系を設定した。

例えば、本品詞体系の品詞がいわゆる学校文法での取り扱いあるいは名称と異なるような箇所には、括弧付で学校文法の品詞を付加しておき、利用者による変更を容易にしている。また、この品詞体系では、「形容動詞」を認めているが、もしこれを「名詞」＋「助動詞」としたい場合には容易に変換が出来る。

「単語」分割においては、どのような単位で分割するかが問題となる。ＲＷＣテキストデータベースの作成に際しては、形態素解析ツールを利用し、構文情報を用いない範囲で処理を行なっているため、形態素単位での分割を行なった。したがって付与する「品詞」も形態素解析レベルで行なえるもの（を中心）とした。

この品詞体系は、ＴＨｉＭＣＯ（Tagset of High quality for Integrated Multi-usage Corpus Openly available to public）と名付けられた。ここでは、品詞は必要に応じて第１レベルから第５レベルまでに詳細化して記述される。第１レベルの品詞としては、以下の１２のものが挙げられている。

（１）名詞　（２）動詞　（３）形容詞　（４）形容動詞　（５）副詞　（６）連体詞
（７）接続詞　（８）助詞　（９）助動詞　（１０）感動詞　（１１）記号　（１２）その他

下位のレベルの分類については言語学的に妥当であり、かつ言語処理に有用と思われるような分類を加えてある。また、人手による修正を前提としているため、本データベースには基本的には、「正解」が記述されている。（人手による修正の段階でも）判断が分かれるような場合には、解釈を保留し曖昧性を残した形の品詞を設定している。たとえば、並立助詞か終助詞かが判断できない「か」に対しては「並立助詞／終助詞」という品詞を与えている。品詞体系の詳細については、参考文献［1］に詳しい。

3.3 データベースの書式

タブコードで区切られた次の項目を１レコードとして持つデータベースを作成する。

（１）分割された単語　（２）読み　（３）原形　（４）品詞分類

実際のデータベースの例を図１に示す。

3.4 作業手順

1. テキストの入手

   通商白書については、平成４年度版および５年度版については、手作業で入力を行なった。平成６年度版およびわが国産業の現状については電子化されたファイルを入手した。しかしながら、このファイルは電子化された最終稿ではあるが、出版物とは微妙に異なる点があり、比較修正した。電子協の報告書については、電子化されたものを入手した。

2. 前処理

   （ａ）一文単位に分割　（ｂ）文ＩＤの付与　（ｃ）特殊文字等の変更

3. 形態素解析

   日本ＩＢＭの形態素解析ツールＪＭＡを用いて自動解析した。ＲＷＣが採用した品詞体系がＪＭＡが元来用いていたものとは異なるため、後処理フィルターを作成し、品詞の置きかえを行なっている。

4. 人間による修正

修正作業は日本語学専攻の大学院生および学部学生が行なった。品詞情報付きテキストデータベースを作成するための援助ツールとして、Nemacs 上で動く（emacs-lisp で記述された）編集ツールを作成した。これは（1）レコードの分割と作成、（2）レコードの結合と削除、（3）品詞選択、の各機能を持つものである。これは上記の書式に基づくデータベースを対象にするツールであり公開可能である。

4 おわりに

実世界の自然言語の解析に関する研究と、その評価に用いるためのテキストデータベースの作成を平成6年度より開始した。これまでに作成したものは1万文強であり、テキストデータとしては分量的には極めて少なく全く不十分なものであるが、今後継続的に拡張していく予定である。

参考文献

[1] 橋本 三奈子、荻野 紫穂、徳永 健伸、元吉 文男、井佐原 均：ＩＰＡコーパスの概要、ＩＰＡシンポジウム'９５論文集（１９９５）

[2] 自然言語処理技術の動向に関する調査報告書、（社）日本電子工業振興協会自然言語処理技術委員会（１９９５）

ＷＧ０６：ｇｅｎｊｏｕ：０００００６０文ＩＤ

なお	ナオ	なお	接続詞			
、	、	、	記号			
最近	サイキン	最近	名詞	副詞可能		
の	ノ	の	助詞	格助詞		
動向	ドウコウ	動向	名詞			
を	ヲ	を	助詞	格助詞		
見る	ミル	見る	動詞	一段	見出し形	
と	ト	と	助詞	接続助詞		
、	、	、	記号			
（中略）						
ようやく	ヨウヤク	ようやく	副詞	助詞類接続		
産業	サンギョウ	産業	名詞			
活動	カツドウ	活動	名詞			
に	ニ	に	助詞	格助詞		
動き	ウゴキ	動き	名詞			
が	ガ	が	助詞	格助詞		
出	デ	出る	動詞	一段	連用タイ接続	
始め	ハジメ	始める	動詞	一段	連用タ接続	非自立
て	テ	て		助詞	接続助詞	
いる	イル	いる	動詞	一段	見出し形	非自立
。	。	。	記号			

図1　実際のデータの例